AUTOMATIC AGE ESTIMATION FROM REAL-WORLD AND WILD FACE IMAGES BY USING DEEP NEURAL NETWORKS

Zakariya Qawaqneh

Under the Supervision of Dr. Buket D. Barkana

DISSERTATION

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE

AND ENGINEERING

THE SCHOOL OF ENGINEERING

UNIVERSITY OF BRIDGEPORT

CONNECTICUT

November, 2017

AUTOMATIC AGE ESTIMATION FROM REAL-WORLD AND WILD FACE IMAGES BY USING DEEP NEURAL NETWORKS

Zakariya Qawaqneh

Under the Supervision of Dr. Buket D. Barkana

Approvals

Committee Members

Name	Signature	Date
Dr. Buket D. Barkana		11/3/2017
Dr. Xingguo Xiong		11/3/2017
Dr. Navarun Gupta		11/3/17
Dr. Miad Faezipour		11,3,2017
Dr. Amalia Rusu		11/3/17

Ph.D. Program Coordinator

Dr. Khaled M. Elleithy		12/13/17

Chairman, Computer Science and Engineering Department

Dr. Ausif Mahmood		12-13-2017

Dean, School of Engineering

Dr. Tarek M. Sobh		12-14-2017

AUTOMATIC AGE ESTIMATION FROM REAL-WORLD AND WILD FACE IMAGES BY USING DEEP NEURAL NETWORKS

AUTOMATIC AGE ESTIMATION FROM REAL-WORLD AND WILD FACE IMAGES BY USING DEEP NEURAL NETWORKS

ABSTRACT

Automatic age estimation from real-world and wild face images is a challenging task and has an increasing importance due to its wide range of applications in current and future lifestyles. As a result of increasing age specific human-computer interactions, it is expected that computerized systems should be capable of estimating the age from face images and respond accordingly. Over the past decade, many research studies have been conducted on automatic age estimation from face images.

In this research, new approaches for enhancing age classification of a person from face images based on deep neural networks (DNNs) are proposed. The work shows that pre-trained CNNs which were trained on large benchmarks for different purposes can be retrained and fine-tuned for age estimation from unconstrained face images. Furthermore, an algorithm to reduce the dimension of the output of the last convolutional layer in pre-trained CNNs to improve the performance is developed. Moreover, two new jointly fine-tuned DNNs frameworks are proposed. The first framework fine-tunes tow DNNs with two different feature sets based on the element-wise summation of their last hidden layer outputs. While the second framework fine-tunes two DNNs based on a new cost function.

For both frameworks, each has two DNNs, the first DNN is trained by using facial appearance features that are extracted by a well-trained model on face recognition, while the second DNN is trained on features that are based on the superpixels depth and their relationships.

Furthermore, a new method for selecting robust features based on the power of DNN and l_{21}-*norm* is proposed. This method is mainly based on a new cost function relating the DNN and the L21 norm in one unified framework. To learn and train this unified framework, the analysis and the proof for the convergence of the new objective function to solve minimization problem are studied. Finally, the performance of the proposed jointly fine-tuned networks and the proposed robust features are used to improve the age estimation from the facial images. The facial features concatenated with their corresponding robust features are fed to the first part of both networks and the superpixels features concatenated with their robust features are fed to the second part of the network.

Experimental results on a public database show the effectiveness of the proposed methods and achieved the state-of-art performance on a public database.

ACKNOWLEDGEMENTS

Completion of this doctoral dissertation was possible with the support of several people. I would like to express my sincere gratitude to all of them. I would like to express my special appreciation and thanks to my advisor Dr. Buket D. Barkana, she has been a tremendous mentor for me. I would like to thank her for encouraging my research and for allowing me to grow as a research scientist.

I would also like to thank my committee members, Dr. Navarun Gupta, Dr. Xingguo Xiong, Dr. Miad Faezipour, and Dr. Amalia Rusu for serving as my committee members even at hardship. I also want to thank you for letting my defense be an enjoyable moment, and for your brilliant comments and suggestions, thanks to you. I am grateful for Dr. Khaled Elleithy, the director of the Ph.D. program, for the academic support and the facilities provided to carry out the research work at the Institute.

Nobody has been more important to me in the pursuit of my Ph.D. degree than the members of my family. I owe everything good in my life to my parents, Mohammad and Aysheh, and for all of the sacrifices that they have made on my behalf. Your prayer for me was what sustained me thus far, your love and guidance are with me in whatever I pursue. Most importantly, I wish to thank my loving and supportive wife, Nayruz, and my two wonderful children, Raya, and Kenan, who provided unending inspiration. Words cannot express how grateful I am to my brothers and sisters, especially my brother Kamel, who has supported me more than I could ever give him credit for here.

A great warm word for my dear friend and fellow researcher Arafat Abu Mallouh, Arafat's continuous support, deep discussions, and persuasive analytical thinking have

enriched our research and always pushed our goals to maximum limits. Arafat was always there during the happy and hard times. For all these reasons and many, many more, I am eternally grateful. Thank you Arafat for making the long journey of pursuing our Ph.D. degrees bearable.

I would like to thank all the staff of the School of Engineering for their support that made my study in the University of Bridgeport a wonderful and exciting experience.

Above all, I owe it all to Almighty God for granting me the wisdom, health and strength to undertake this research task and enabling me to its completion.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AAM	Active Appearance Model
AGES	Aging Pattern Subspace
ANN	Artificial Neural Network
BIF	Bio-Inspired Features
CNN	Convolutional Neural Network
CS	Cumulative Score
DNN	Deep Neural Network
HCI	Human-Computer Interaction
ICA	Independent Component Analysis
LBP	Local Binary Patterns
LDA	Linear Discriminant Analysis
LLE	Local Linear Embedding
MAE	Mean Absolute Error
MSE	Mean Squared Error
PCA	Principle Component Analysis
SFP	Spatial Flexible Patches

CHAPTER 1: INTRODUCTION

The human face contains various information such as age, gender, emotional state, pose, and ethnic background. Such information could be extracted and used in entertainment, cosmetology, biometrics, human-computer interaction (HCI), security control, and surveillance monitoring applications. Recent developments in computer technology have a direct impact on the growth of the image processing techniques while enriching the applications in computer vision and graphics fields further. One of the most popular research fields which have gained attention over the years is the automatic estimation of age information from facial images. Age estimation is defined as determining the exact age or the age range of a person using 2D facial images [1, 2]. Age classification from facial image consists of two main parts. The first part is the prepossessing and the feature extraction. While the second part is the classification process. In the first part, some techniques are applied for processing the image from many challenges such as low-resolutions, various expressions, and occlusions. Then features are extracted from the images. These extracted features are fed for a classifier and it is trained to find a pattern for each age class to distinguish the classes from each other. The global spreading of the internet, smart phones, and social media application created a relatively new market of online services that depends on facial biometrics for performing a wide variety of services. The later imposed an urgent need for accurate and trust worthy applications that can extract and classify the required biometric. Currently, most of the captured facial images are described as unconstrained, there are no prior conditions on the place, illumination, background, or the pose. The challenge here is to find a distinguished feature set and an

efficient classifier that both can perform a recognition task from such variant and unconstrained images. Age classification based on facial images has several applications in different domains and it is also expected to involve more applications in the near future. The importance of having reliable age classification systems is related to the type of application attached to the task. For example, determining the type and dose of a medicine depends on the age of the patient, as a result, if the age of the patient is to be estimated using a computerized system then an accurate and reliable age classification system is a must.

In this research, in order to enhance the age classification from the facial images, it is planned to extract several distinctive feature sets that contains more age-related information. A DNN classifier is used since it is nowadays one of the most successful that achieved the state of the art of several classification and estimation fields. It is aimed to enhance the classification processes by introducing different architectures and ways.

1.1 Main Facial Image Related Fields

Recently finding a semantic information from images has increased and has gained the attention of the research studies due to the huge number of images which are added daily on the internet or being stored on personal phones and computers. To extract such semantic information efficiently and automatically, these images should be analyzed and indexed in very organized method using different machine learning techniques. Especially facial images contain various valuable semantic information about the human being. This information might be used in different fields such as face recognition, facial expression

recognition, emotion recognition, head pose and gaze recognition, and age and gender recognition.

1.1.1 Face recognition

Over the last decade face recognition has received substantial attention in diverse image analyses fields such as in biometrics [3-4], computer vision [5], pattern recognition [6-8], and computer graphics [9-10]. As well as, nowadays face recognition technique is used in several of forensic, security and commercial real applications. Face recognition consists of two main subfields: Face verification and face identification.

- Face verification: verification system tries to verify the identity of the claimed person or the claimed biometric of the person. In other words, the system tries to answer the question "Am I who I say I am?" for Several algorithms and techniques have been proposed. Several studied have been conducted for face verification as in [11-13].

- Face identification: In face identification, the system tries to identify unknown person or unknown biometric. Therefore, the system tries to answer the question "Who am I?". Several studied have been conducted for face identification as in [14-16].

Several methods have been developed and introduced for face recognition these methods can be categorized into appearance based models and model based model.

1.1.1.1 Appearance-based Models:

In appearance model the face image is recognized as high dimensional vectors of features. These vectors are used to drive and efficient and effective representation for the image in the vector space. In testing phase, the similarity between the stored model for each face and the test face image is found in the feature space. Appearance models proposed to represent the face image are classified into two main categories:

- Linear models: Typically, all these models statically try to represent each image vectors with several coefficients by projecting these vectors to the basis vectors. Several techniques have been proposed for linearly analyses the face images such as PCA [17], ICA [18], and LDA [19-20].

- Non-Linear (manifold) models: Linear models are considered as a simple approximation of the non-linear models; therefore, the manifold models are more complicated than the linear models. In non-linear models, a nonlinear mapping from the input space is derived to find the most significant feature into the feature space. In the past, several methods have been proposed such as kernel-PCA [21], ISOMAP [22], and LLE [23]. Recently deep learning is widely used as non-linear model for face recognition. Several algorithms based on deep learning have been proposed such as in [24-29].

1.1.1.2 Model-based:

This model tries to construct a model for each face that capture the facial variations from the available facial images for each person. Model based proposed to represent the face image are classified into two main categories:

- 2D models: One of the most famous model 2D model is active appearance model (AAM). AAM model combined a representation of a statistical model of the face shape and the gray level appearance of the face.

- 3D models: In 3D space, the model is built with care of different facial variations such as pose and illumination, therefore, representing the facial image based on the 3D technique is more efficient and better than representing the face as 2D model. For face recognition, a 3D model that represent the shape and the texture in terms of model parameters is used and separated from the extrinsic model parameters of the face such as pose and illumination.

1.1.2 Facial expression recognition

This field studies the changes in facial expressions that results from different factors such as intentions, communication, and emotional state [30-31]. Facial images are the main source for recognizing the changes in the facial expressions by using an automated. Facial expression recognition has a wide range of possible applications, for example, clinical psychology, lie detection, HCI, and pain assessment. In the literature, different studies have been carried out for developing new techniques for better facial expression recognition as in [32-37].

1.1.3 Emotion recognition

Emotion recognition needs higher level of information than facial expression recognition and it plays a major role in human-human and human-machine communication [38]. An efficient emotion recognition system should be able to differentiate between different facial expressions to classify them into different categories of emotions. The

emotion categories vary from simple ones like happy and sad to complicated ones such as guilty, worried, and bored. This field have many challenges resulted from the quality of the input facial images and the cultural background of the subjects [39]. HCI is one of the major applications that depends on a successful facial emotion recognition system. Many studies explored this field with different degrees of success as in [40-45].

1.1.4 Head pose and gaze recognition

This part is a major field used in HCI where both information about the pose of the head and the eye gaze are used to determine the subject's point of attention accurately [46-48]. Being able to estimate the head pose is considered to be an important step before performing other tasks such as face detection and recognition. Another important application where head pose and eye gaze are considered as a necessary pre-step is the 3D modeling of facial pictures. The accurate estimation of the head pose and eye gaze has a major effect on the resulted 3D model. Currently, new methods for recognizing the head pose and eyes are being utilized for new trends in HCI and in the widely used smart phones market [49-53].

1.1.5 Gender recognition

Recognizing the gender of a human from a 2D image has different applications. It has been studied earlier in the last two decades and continued to be the focus of many research groups due to the rapid development and spread of communication devices such as smart phones and smart TVs. The new smart phones are equipped with accurate cameras which allowed the users to capture images almost everywhere and on any time. There are several applications for automatic gender recognition such as visual surveillance, smart

marketing systems, and security control [54-55]. In the literature, two types of images were used for gender recognition, constrained images and unconstrained images. Constrained images are captured under controlled environments where the pose and the illumination are controlled before capturing the image. While unconstrained images are captured in variant and uncontrolled environments. Clearly, unconstrained images are more challenging for gender recognition. Different features were used for performing successful gender recognition, for example, raw pixels, Haar-like features, LBP features, and fragment-based filter banks [56-59]. Many methods were used for achieving reliable and accurate gender recognition such as Anthropometric Models and appearance models, and recently, deep neural networks methods have shown remarkable results [60-63].

1.2 Motivation behind the Research

The classification of age from 2D images gains importance in many present and future applications including education, criminal cases, advertisement, phone ads, merchandise, controlled media access, statistics of population, electronic health applications, information forensics and security [64] and more. At present, age appropriate education, ads, and merchandises can be offered to users [65]. Media content can be made available based on the user's age [66]. Research in artificial intelligence is rapidly developing. In near future, a computerized system called robot doctors may determine the correct dose of medicine for a patient depending on his/her age. Smart robots may select the right age appropriate attitude and language while socializing with humans. There are several challenges in age estimation. One of the main challenges is that people do age at variable rates that are affected by factors such as, genetic factors, social conditions, and

life style. Some people can look years younger than their chronological age while some can look years older. Another challenge is the dissimilarity between aging rates of men and women. Wearing makeup and accessories either to look younger or to hide aging marks is another challenge [67]. Age classification from facial images is still an open research area. Although, there are some researches have been conducted to enhance the age and gender from facial images, they did not achieve good accuracy results especially for the age from unconstrained facial images.

With these challenges are in mind, in this research it is aimed to find and extract more distinctive features that contains information related for the person age from unconstrained 2D images based on CNNs and DNNs while utilizing different DNNs architecture for improving the classification processes for age classification from facial images problem. Recently, CNNs and CNNS have shown remarkable performance in various computer vision fields, such as object recognition, face detection, and human pose estimation.

Existing benchmarks for age classifications are relatively small compared to the benchmarks used in face recognition. Training a deep neural architecture using a small benchmark is problematic since training a very deep CNN architecture on a relatively small benchmark is liable to a critical overfitting. Therefore, deep CNN architectures that are trained for other classification tasks such as image classification, semantic segmentation, and face recognition on large benchmarks are employed. Then the experimental results are analyzed about how these pre-trained models can be used to find more representative features and how they can be adapted and fine-tuned to estimate the age of a subject from an unconstrained 2D image.

1.3 Main Contributions of the Proposed Research

The focus of this research is to enhance age classification based on unconstrained 2D images. To achieve this goal, this research addresses the problem from different angels as follows:

- Finding a distinctive feature set that contains specific and accurate information about the subject's age depending only on the available 2D images of the subject. It is proposed to use pre-trained CNNs for different task on large benchmarks to extract facial features that will perform well for age estimation successfully and it is shown show that features extracted from pre-trained models for domain specific tasks can be successfully used to improve the age classification task.
- We find a new feature set based on the superpixels depth and their relation. The image will be divided into a number of small regions called superpixels where each superpixel represents a group of pixels. It is assumed that the depth contains age-related information.
- To enhance the classification process, it is proposed to jointly fine-tune two DNNs with different feature sets. The first DNN is trained by using the first feature set. The second DNN is trained by the second feature set. Then, their last hidden layer outputs are element-wise summed to be trained and fine-tuned jointly.
- Developing a new cost function that can calculates the error during the training process for a big number of examples, on condition that, the learning process will be optimized and converge with lesser overfitting effect.

- Introducing a new architecture for the classification process that jointly fine-tuned two different DNN architectures based on the new proposed cost function and feature sets.
- Selecting robust feature set for age classification based on the power of $l_{2,1}$-norm and DNN using a new cost function.

CHAPTER 2: LITERATURE SURVEY

Over the last ten years, many studies have been carried out on the age estimation from real-world and wild facial images. In this section, existing benchmarks are presented and a brief review of the most significant and milestone works is given with regard to feature extraction and classification methods.

Kwon and da Vitoria Lobo [68] carried out one of the earliest works in age estimation by using facial images. Cranio- facial changes in feature-position ratios and skin wrinkles were used as features for three age groups (baby, young adult, and senior adult). Facial features were detected and their ratios were computed. Skin wrinkle analysis was performed. This early work in 1994 has shown that computing ratios and detecting the presence of wrinkles can yield the age from facial images. The same year, Farkas [69] presented a mathematical model to estimate the growth of a person's head from infancy to adulthood. This model was used to estimate the age of a person from a facial image. The drawback of this model is that the performance of the age estimation degrades for adults by using models which are built by using 2D images. One of the earliest research works in age estimation is based on face anthropometry. Face anthropometry is a science that deals with measuring sizes and proportions on human face. In general, the estimation of age from facial images using anthropometry model is limited to young ages. The shape of the human head does not change significantly in the adult years. Moreover, the ratio of distances for face geometry is calculated using 2D images but 2D images are sensitive to head pose. As a result, frontal face images are the only images that can be used to measure geometry of

the face. Therefore, anthropometry is not suitable for age estimation by using real-world and wild facial images.

Other approaches have been proposed based on the facial features (appearance model or face descriptor). Local and global facial features were extracted [70-71]. Texture and shape features were calculated by using a semantic-level description of the face to describe facial features. They also built a classification system to estimate different age groups with five year intervals. Their system was tested on a Japanese database of 500 subjects aging from 15 to 64 years old. Moreover, gender estimation was done to improve the performance of the age estimation. Since women and men have different aging rates, the inclusion of gender estimation enhanced the age estimation.

Ramanthan and Chellappa [72] worked on age progression in young face images and computed eight ratios of distance measures for modeling age progression. They proposed a craniofacial growth model by illustrating how the age-based anthropometric constraints on facial proportions translate into linear and non-linear constraints on facial growth parameters and proposed methods to compute the optimal growth parameters. The purpose of their work was to predict one's appearance across the years and to perform face recognition. Anthropometrical changes of human face and its size, shape, and textural patterns may adequate to estimate an individual's age up to the adult years.

Lanitis et al. [73] studied the aging effects on face images and described how the effects of aging on facial appearance can be explained. They built a statistical-based face model. By the proposed shape intensity face model and automatic age simulation,

statistically significant improvement in the performance of the age classification system was reported.

Geng et al. [74-75] modeled the aging pattern as the sequence of an individual's face images sorted in time order by constructing a representative subspace. The AGES model built an aging pattern for different age stages. In case, the images of some ages were not available, they were synthesized by using EM-like iterative learning algorithm. The AGES was evaluated on the FG-NET database with a mean absolute error of 6.77 years. They reported that the performance of the model was significantly better than the existing age estimation methods in 2007 and was comparable to that of the human observers. One of the limitations of AGES approach is that it assumes the availability of images representing the different ages of an individual. If images for different ages are not available, AGES assumes that there is an age pattern similar to the input image. AGES approach utilizes the AAM to calculate the face representation to encode the wrinkles of the face. AAM only encodes the image intensities which cannot describe the local texture information. Local texture information is important to represent the wrinkles of elderly people.

Manifold analysis is used and proved to be promising in age estimation from face images by several studies [76-78]. An age estimation framework was proposed by Fu et al. [76] using manifold analysis and learning methods to find a sufficient low-dimensional embedding space. Manifold data points were modeled with a multiple linear regression function. Age manifold model is more flexible than AGES, since images could be built using different person's images for unavailable images of some ages. Age manifold built the common aging pattern using the manifold embedding technique to learn the low

dimensional aging trend from a group of face images for each age. Scherbaum et al. [77] also proposed a statistical age estimation method using manifold learning over a 3D morphable model.

Gunay and Nabiyev [79] used effective texture descriptor for appearance feature extraction and utilized LBP in automatic age estimation system. Using the nearest neighbor classification, their system achieved 80% accuracy on the FERET database. By using AdaBoost, they achieved 80-90% of accuracy on the FERET and PIE databases. Gao and Ai [80] used the Gabor feature with fuzzy-LDA for age estimation. Their work showed that Gabor feature is more effective than LBP. Yan et al. [81-82] employed SFP to be the feature descriptor in order to handle images with small undesirable defects such as occlusions and head pose. They achieved MAE accuracy of 4.94 years on the FG-NET database. Sparse feature design, graphical facial features topology, geometry, and configuration, were proposed and age estimated based on the multiresolution hierarchical face model by ANN [83]. This system achieved MAE of 5.974 years on the FG-NET database. In [84], Mu et al. proposed BIF for age estimation. The bio-inspired features have the ability to handle small rotations and scale changes effectively. By using the BIF with an SVM classifier, their work achieved MAE of 4.77 years on the FG-NET database. In [85], two feature sets, BIF and the age manifold were used. By using an SVM classifier, the system achieved MAEs of 2.61 and 2.58 years for female and male on the YGA database, respectively.

Literature has developed over time to enhance the performance of age estimation from face images. Naturally, each method tried to overcome the limitations of the previous methods by widening the range of domains. All the previously mentioned methods showed

conditional significant performance where the used databases were either small on size or constrained to specific kind of images such as frontal and aligned pose images. On the other hand, the proposed work is applied on a larger database with unconstraint face images.

In [86], age estimation on real-life faces acquired in unconstrained conditions was studied. The LBP and Gabor features were exploited as face representation. Adaboost was used to learn the discriminative LBP-Histogram bins for age estimation. They achieved 55.9% of accuracy on the Group Photos benchmark by an SVM classifier. Alnajar et al. [87] adopted a learning-based encoding method for age estimation under unconstrained imaging conditions. Multiple codebooks for individual face patches were extracted and learnt. The orientation histogram of local gradients as the feature vector for code learning was used. An unconstrained database Group Photos benchmark which contains 2744 images was used and they achieved an absolute improvement of 3.6% over the study in [86] on the same database.

In [88] they proposed a framework for estimating the age, gender, and ethnicity jointly. They investigated different techniques for achieving better results. They utilized the linear and nonlinear canonical correlation analysis and the partial least squares models using their joint framework. Their analysis was conducted based on the rank theory. They showed that the bio inspired features could be used to represent the face image for the three labels, age, gender, and ethnicity. They evaluated their work on the MORPH database, for age estimation they achieved 3.98 MAE using regularized kernel canonical correlation analysis. When the support vector machine or support vector regression are used for

classification with the features extracted by canonical correlation analysis and partial least squares the MAE on MORPH dataset is decreased to 3.92.

In [89] they proposed a cost sensitive local binary feature learning for age estimation using facial images. They reported that their method represents the face from image pixels using discriminative local features. In their work, from the face patches they extracted low-dimensional binary codes from the raw pixels using several hashing functions. Also, real valued histogram features were calculated from the binary codes to represent the face. In addition, to learn the hashing functions jointly they proposed a cost sensitive local binary multi feature learning. For evaluation, they tested their work on FG-NET, MORPH, LifeSpan, and FACES datasets. They achieved 4.36 MAE on FG-NET, 4.37 MAE on MORPH, 5.26 MAE on LIFESPAN neutral faces, and 4.84 MAE on FACES neutral faces. The advantage of their work is the ability to learn the features directly from the raw data, also they stated that using the binary information is better because it is not affected with local variations.

Recently, new benchmarks have been designed for the task of age estimation from face images. These new benchmarks are more challenging than the previous benchmarks in terms of quantity and quality. The size of the new benchmarks is much larger and most importantly the quality of the included images is categorized as unconstrained. The unconstrained images reflect the real world wild environments and are collected from online image repositories. The Group Photos [66] and the Adience benchmarks [64] are examples of these new benchmarks. Currently, the Adience benchmark is considered to be the newest and the most challenging benchmark for age and gender estimation from face images.

Recently CNNs and DNNs have been started to use for age estimation from face images. In [67], a simple CNN architecture was used as a feature extractor and a classifier to avoid overfitting problem. They evaluated their work on the Adience benchmark and achieved 50.7% overall accuracy. [90-91] proposed new systems for age regression based on facial identification features by using a relatively small deep CNN model. Both works used the same face identification model in [92] for feature extraction. In [90], a cascaded classification and regression system based on a coarse age classifier has been proposed. They introduced an age regressor for each age group based on the features extracted from the coarse age classifier. Then they used an error correcting method for correcting the regression error for subjects. [91] proposed a system that the features were extracted from a pre-trained CNN for face identification. The extracted features were fed to a small neural network to regress the age of the subject.

[93] proposed two approaches for age estimation, the first approach is the fusion of descriptors based on texture and local appearance. In this approach, they combined well-known local descriptors that can capture texture and contour cues, they reported that using texture and contour cues together enhanced the performance of age estimation better than using each feature alone. They used the histograms of oriented gradients, the local binary patterns, and the speed-up robust features as descriptors. For classification, they used and modified the canonical correlation analysis to find the optimal weights between the data and their labels. The second approach is a deep learning scheme for accurate age estimation. They used convolutional and pooling layers followed by fully connected layers to globally interrelate features. To evaluate their ideas, they tested their work on the MORPH and FRGC databases, where the mean average error and the CS metrics were used

for measuring the performance. Using their first approach, the best results for the MORPH database was 4.25 MAE and 71.2% CS. For the FRGC database, the best results were 4.17 MAE and 76.2% CS. They achieved 3.88 MAE on the MORPH database and 3.31 MAE on the FRGC by using their second approach.

[94] proposed a generic deep network model for automatic age estimation. Their model extracts facial features using a convolutional scattering network, then the dimension of these features is reduced using PCA. They reported that the scattering features are discriminative and invariant translations. The last step is to estimate the age using three fully connected layers that act as category-wise rankers. They used the rank value to investigate the relation between the age labels, also the category rankers estimate age within the class. For evaluation, they tested their work on the MORPH, Lifespan, and FACES datasets. The results were 3.49, 5.19, and 7.04 MAEs respectively.

In [95] a CNN is used to estimate the age from the image. In this work, local aligned patches were extracted using several facial landmarks and each patch feed to different CNN. For each face image, 21 facial landmarks were extracted. Then the landmarks are grouped to 13 pairs. 48 x 48 patches were cropped in 4 scales. In total 23 paths pairs were extracted. Then for age estimation, each patch of the 23 patches was trained in separate DNN to learn each patch features. After training the 23 DNNs, their final fully connected layer outputs were fused to estimate the age of the person. For evaluation of the proposed work MORPH Album 2 was used. And they reported an average of 3.63 MAE.

[96] proposed two methods for age estimation. In the first methods, a distribution-base loss function using CNN was introduced. They stated that the usage of distributions

as labels is better than age labels for training since the distributions can utilize the uncertainty of the manual labeling. They used two CNN architectures where the first CNN was the VGG-16 and it was fine-tuned three times using different datasets, and the second CNN was trained using different types of input data. After that, the two CNNs are fused using the ChaLearn database. The KL divergence loss function was used as a distribution-based function to exploit the uncertainty, also they used the softmax loss function. For fusion, they used the distance-based voting ensemble method to predict the age from concatenated feature resulted from tuning and training over different datasets. In the second method, they proposed a new CNN to be trained on a different type of inputs such as: 1) RGB color-space of the aligned face image, 2) Image gradient magnitude and orientation using the gray-level, 3) The HSV color-space images, and then fine-tuned on the Chalearn data set. For training and testing, they used the ChaLearn data set and a collection of images from different data sources such as MORPH, Google image search, FG-NET, and Adience. They achieved 0.305 MAE over the ChaLearn test set.

In [97] AgeNet was proposed to estimate the age apparent for the ChaLearn 2015 Apparent Age database. Two different CNN models were trained and fused to estimate the apparent age. The first kind models are based on real-value regression models, while the second kind models are classification models based on a Gaussian label distribution. Both models used features extracted from pre-trained models for face identification. To reduce the risk of over-fitting, the AgeNet is pre-trained on large database for face identification, since the age is estimated from face images it is expected that the feature extracted for face identification is correlated to the age estimation problem. The CASIAWebFace is used to train the AgenNet for face identification. After that the AgeNet was pre-trained to estimate

a real age from a face images using 3 different databases, including CACS, Morph-II and WebFaceAge. Then the AgeNet was fine-tuned to be trained on the training set of the ChaLearn 2015 Apparent Age database for evaluation and test. The AgeNet was achieved 3.3345 MAE for the apparent age on the Chalearn 2015 database.

[98] Introduced the IMDB-WIKI dataset which is one of the largest datasets with labeled age and gender images. Also, they proposed to solve the real and apparent age estimation problem based on deep learned models from large data, robust face alignment, and expected value refinement after formulation of age regression value. For the face alignment, they proposed a new technique that depends on the rotating the image over different angels and then running the face detector on the original and rotated images to select the image with the highest detection score. The pre-trained CNN was trained for age classification for the training data set, where the age values are distributed into ranges of age. Each range covered continuous values of ages. In the test phase, the expected value over the softmax-normalized output probabilities of the age ranges was computed to represent the estimated age for the input image. For evaluation, they tested their work on different datasets such as, ChaLearn, MORPH, FG-NET. The best result achieved for testing over ChaLearn dataset for apparent age was 3.252. For real age estimation, the proposed work was tested over MORPH and FG-NET datasets. The results were 2.68 MAE on the MORH and 3.09 on the FG-NET.

CHAPTER 3: PROPOSED FEATURE SET AND METHODS BASED ON DNN

This section describes our strategies and methods which are proposed to enhance the age classification problem. Several approaches have been proposed and analyzed. The potential problems for each method have been discussed and alternative strategies for solving the problems have been introduced. The rest of this chapter is organized as following: Section 5.1 explains our proposed method for utilizing pre-trained CNN models for feature extraction and classification. Section 5.2 introduces the proposed Jointly Fine-Tuning DNNs using different feature sets for age classification.

3.1 Utilizing Pre-trained CNN Models for Feature Extraction and Classification

3.1.1 Efficient feature sets extraction and classification

Motivated by the success of CNNs architectures in different fields, CNNs are used as feature extractors for automatic age estimation. Existing benchmarks for age classifications are relatively small compared to the benchmarks used for other classification tasks such as image classification, semantic segmentation, and face recognition on large benchmarks. Training a deep neural architecture using a small benchmark is problematic since training a very deep CNN architecture on a relatively small benchmark is liable to a critical overfitting. To overcome this problem, deep CNN architectures that are trained for other classification tasks are employed. Then, these architectures are adapted and fine-

tuned to estimate the age of a subject from an unconstrained 2D image. The idea here is to take advantage of a large database and the deep architecture of the network, which is designed on a large database. Deep network architectures trained on large databases are capable to extract distinctive and robust features and they are less prone to overfitting. Building deep network architecture for age estimation on small databases is expected to have poor performance.

3.1.1.1 Architecture for feature extraction and age classification

In this section, our proposed architecture for fine-tuning pre-trained models for age classification will be explained by using a model trained for face recognition task as an example. Using an efficient facial feature extractor is expected to perform well for age estimation. However, a trained face recognition model that was trained on a large database may extract facial features more efficiently than training a new model on a small database. The idea here is to train a deep network to study facial features from image and then retrain and fine-tune this network to estimate the age information.

In this work, the CNN architecture proposed by [99] is considered. It achieved comparable results to the state-of-the-art for face recognition task (VGG-Face). In [99], three CNN architectures named A, B, and C were used. The architecture A is used. Details of the large database and the configuration of the architecture A can be found in [99]. The architecture A consists of eight convolutional layers and three fully connected layers. A rectification operator is used after each convolutional operator. A max pool operator is added at the end of each convolutional layer. 4096-dimensional output is used for the first two fully connected layers. Dropout with p= 0.5 and rectification operator are applied to

the first two layers. N-way class prediction is used for optimizing the network parameters. Therefore, the output size of the last layer is chosen to be 2622, which represents the number of subjects in the large database.

Here, the convolutional layers of the CNN VGG-Face are reused, while the fully connected layers are replaced with new layers. The modification of the CNN is performed by removing the fully connected layers and replacing them with four new fully connected layers of different sizes as shown in Figure 1. The sizes of the first three fully connected layers are chosen as 4096, 5000, and 5000, respectively. Each of which is followed by two manipulation layers, one dropout layer, and one normalization layer. The last fully connected layer is a softmax layer with a size of 8, which represents the number of labels in the Adience database. Each label represents an age range. The probability of each label is used to estimate the age of corresponding face image. The eight convolutional layers which were used during the training for face recognition task are reused in the modified CNN as shown in Figure 1.

The details of the proposed network architecture and configuration are given in Table 1. The weights between the new layers are initialized by a Gaussian distribution with zero mean and 10^{-2} standard deviation. The new network is trained only for the newly added fully connected layers while keeping the original convolutional layers frozen during the training. This approach appears to be very fast since only the newly added fully connected layers are trained.

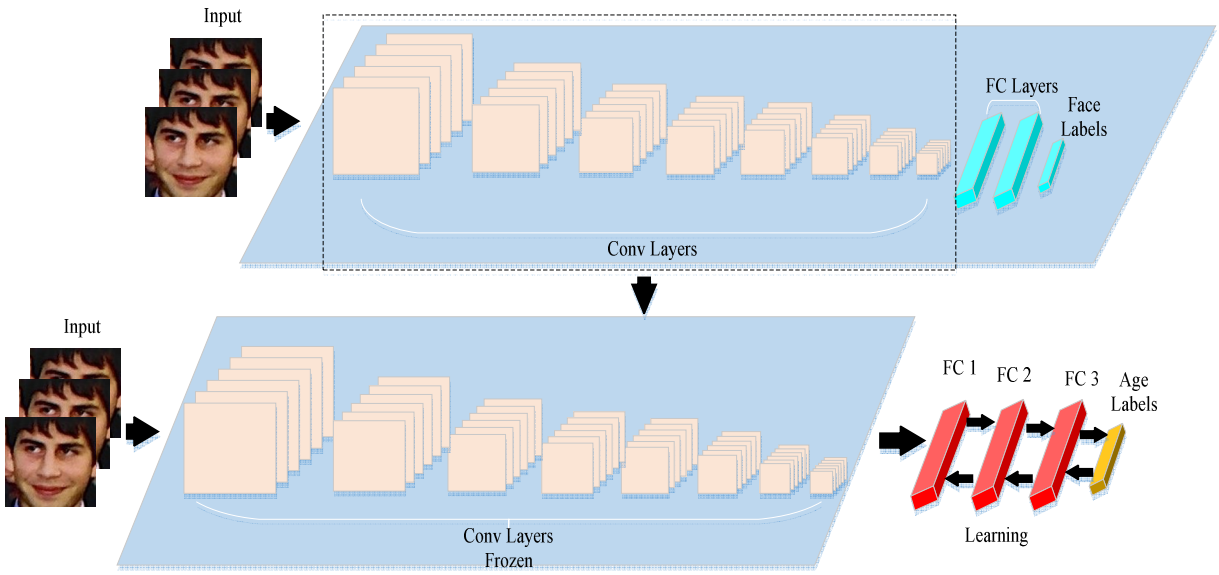

Figure 1. Fine-tuning domain specific pre-trained model for age classification.

Table 1. Network architecture and configuration. For each convolutional layer, number of filters, the filter size, their receptive field convolution stride, and spatial padding are indicated.

layer	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
type	Input	Conv	Relu	norm	mpool	conv	Relu	norm	mpool	conv	Relu	Conv	relu	conv	Relu
name	n/a	conv1	relu1	norm1	pool1	conv2	relu2	norm2	pool2	conv3	relu3	conv4	relu4	conv5	relu3_2
support	n/a	3	1	3	1	2	3	1	3	1	2	3	1	3	1
filt dim	n/a	3	n/a	64	n/a	n/a	64	n/a	128	n/a	n/a	128	n/a	256	n/a
num filts	n/a	64	n/a	64	n/a	n/a	128	n/a	128	n/a	n/a	256	n/a	256	n/a
stride	n/a	1	1	1	1	2	1	1	1	1	2	1	1	1	1
pad	n/a	1	0	1	0	0	1	0	1	0	0	1	0	1	0
layer	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
type	conv	Relu	Mpool	conv	relu	conv	Relu	conv	relu	mpool	Conv	Relu	conv	relu	Conv
name	conv3_3	relu3_3	pool3	conv4_1	relu4_1	conv4_2	relu4_2	conv4_3	relu4_3	pool4	conv5_1	relu5_1	conv5_2	relu5_2	conv5_3
support	3	1	2	3	1	3	1	3	1	2	3	1	3	1	3
filt dim	256	n/a	n/a	256	n/a	512	n/a	512	n/a	n/a	512	n/a	512	n/a	512
num filts	256	n/a	n/a	512	n/a	512	n/a	512	n/a	n/a	512	n/a	512	n/a	512
stride	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1
pad	1	0	0	1	0	1	0	1	0	0	1	0	1	0	1
layer	30	31	32	33	34	35	36	37	38	39	40	41	42		
type	relu	Mpool	Conv	relu	dropout	conv	Relu	dropout	conv	relu	Dropout	Conv	softmx		
name	relu5_3	pool5	fc6	relu6	drop6	fc7	relu7	drop7	fc8	relu8	drop8	fc8	prob		
support	1	2	7	1	1	1	1	1	1	1	1	1	1		
filt dim	n/a	n/a	512	n/a	n/a	4096	n/a	n/a	5000	n/a	n/a	5000	n/a		
num filts	n/a	n/a	4096	n/a	n/a	5000	n/a	n/a	5000	n/a	n/a	8	n/a		
stride	1	2	1	1	1	1	1	1	1	1	1	1	1		
pad	0	0	0	0	0	0	0	0	0	0	0	0	0		

3.1.1.2 Training for age classification

The stochastic gradient descent algorithm is used to train the modified CNN in order to find the optimal parameters which will enable the network to achieve better classification accuracies. 224x224 pixel scaled images are used. The output of each layer is forwarded to the next layer as an input until the softmax layer calculates the probability of each label. The learning is performed only on the fully connected layers. It means that the parameters of the convolutional layers are frozen, while the parameters of the fully connected layers are allowed to be changed. It is shown in Figure 1. Freezing the training on the convolutional layers ensures that the process of extracting facial features is unchanged. The learning rate is set initially to 0.1 and then decreased by a factor of 10 if there is no improvement in the validation set learning. The dropout value is chosen as 0.6. It is observed that using a weight decay together with dropout technique have a positive effect on the classification accuracies [100]. The weight decay is normally set to 10^{-4} or 10^{-5}. The later value has worked fine for the modified network but increasing the value to 10^{-3} has provided a higher overall accuracy. It is being said, deploying dropout technique together with the weight decay as regularizers has no negative effects on the training of the fully connected layers. On the contrary, increasing the value of the weight decay to 10^{-3} enhances the learning process and the accuracies. Since the convolutional layers are trained previously and training occurs only on the newly added fully connected layers, the weight decay value is increased with an increased network performance. The modified network has faster convergence as it only trains the fully connected layers as oppose to train all layers in the original CNN.

We will utilize five more pre-trained models to extract face-related features. Same training and fine-tuning procedure that were applied on the VGG-Face model will be applied on the five models. These models are retrained, fine-tuned and tested for age classification. The architecture of the fully connected layers used for each CNN is summarized in Table 2. CNNs are architected and fine-tuned to predict the age by changing the fully connected layers and their number nodes. Then, each network is trained and fine-tuned while the convolutional layers are kept frozen during the training as explained earlier.

Table 2. Different CNNs architectures.

Network	# of fully connected layers	# of nodes/layer
GoogLeNet	4	1024, 2048, 2048, 8
ResNet-50	4	2048, 5000, 5000, 8
VGG-VD-16	4	4096, 6000, 6000, 8
VGG-VD-19	4	4096, 6000, 6000, 8
FNC-8s	4	4096, 5000, 5000, 8

The five models are: GoogLeNet [101] and ResNet-50 [102] architectures which performed exceptionally well in ImageNet ILSVRC14; the VGG-VD [103] models with 16 and 19 layers, trained on ImageNet ILSVRC for image classification; and FNC-8s [104] trained for semantic segmentation.

3.1.1.3 Prediction for age classification

A given test image is rescaled to 256 x 256, and then three images of 224 x 224 pixels are extracted. The first image is extracted from the center of the original image. The

second and the third images are cropped from the upper-right and the bottom-left corners of the original image, respectively. Then, the trained network is applied densely on the three images. The softmax probability score vectors of the three images are averaged to obtain a final vector of class scores for the original test image from the three images. This method reduces the impact of the challenges such as low-resolutions, various expressions, and occlusions in the database.

3.1.2 Improving the extracted feature sets by dimensionality reduction

There are many cases where the measured or the observed data vectors are described as a high dimensional data. Normally, a significant portion of the high dimensional data is redundant and has low variance, undesired, or resulted from linear operations over other desired data. The goal of dimensionality reduction is to reduce the dimension of the high dimensional data to a smaller one while preserving the same useful or desired information. There are many benefits of dimensionality reduction, for example, it reduces the space needed to store the data during training, it decreases the time needed to process the data, and it increases the performance of the data in many classification tasks [105-106]. In age and face recognition tasks, the number and the size of the input images are considerably large and require careful processing in order to extract and select the distinctive features. As shown in Figure 2, different deep convolutional models that were trained for different tasks other than age estimation are used. PCA is applied to the last convolutional layer output of these models for dimensionality reduction.

The input images are fed to each deep trained model until the output features of the last convolutional layer are calculated. These features are stacked together for the entire training data set. This results in a high dimensional feature vector.

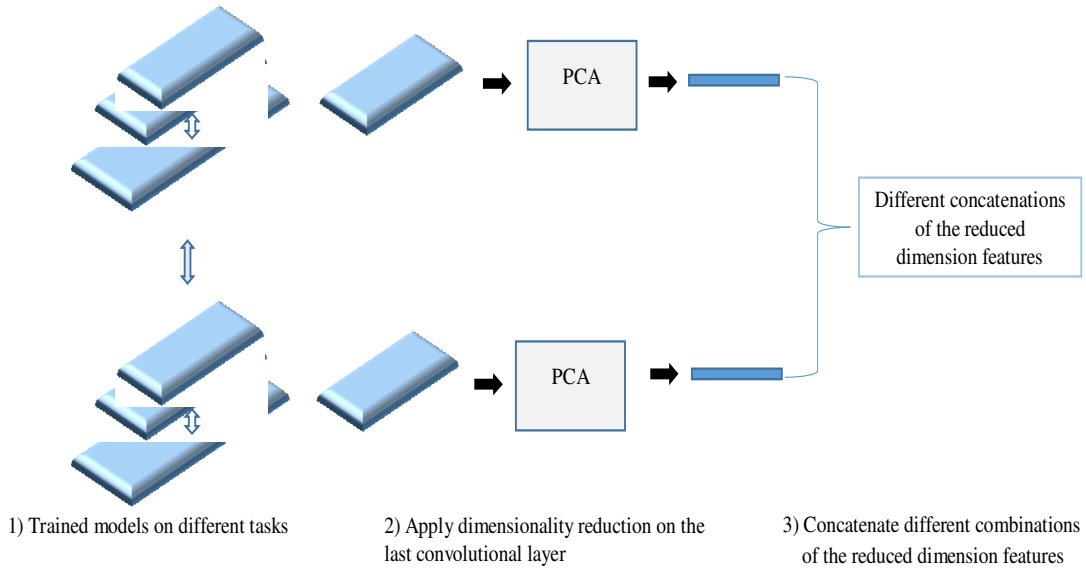

Figure 2. Features dimensionality reduction.

The size of the last convolutional layer differs between the trained models and it is large for all models. As a result, dimensionality reduction is required to fine tune the feature vectors. Since each trained model was trained for a different task, it is expected to have different feature vectors with a different level of performance in age estimation. In this work, the dimensionally reduced features from each trained model and their combination are examined. PCA technique is used for dimensionality reduction. PCA transforms the large space into a smaller subspace using linear transformation.

3.2 Jointly Fine-Tuning DNNs Using Different Feature Sets for Age Classification

In this section, two types of joint fine-tuning architectures are introduced for age classification based on facial images. Each type has different advantages than the other type. It is shown how each type can improve age classification even if used with different feature sets. Section 3.2.1 will introduce the first type which is based on joint fine-tuning based on amplified features, while section 3.2.2 explains the joint fine-tuning of two DNNs based on a new proposed cost function.

3.2.1 Jointly fine-tuning DNNs based on amplified feature sets

In this section, different DNNs are jointly fine-tuned with different feature sets for age classification. The feature sets should be extracted for the training and testing images examples. Any combination of feature sets related to the age classification can be utilized in this method. Two distinctive feature sets are chosen to show the effectiveness of the proposed method in age classification. However, the proposed method can be applied on more than two feature sets. The first feature set is extracted using deep pre-trained model for face recognition. While for the second feature set, the depth features of the facial image based on image superpixels and their relation as new feature set for age classification are proposed. The depth of the image superpixels will reveal aging remarks on different parts of the face and this will help to estimate the age interval of the subject. In addition, the superpixels depth and their relation will add more information to reveal hidden aging remarks.

3.2.1.1 Pre-trained CNN Model-I

The pre-trained CNN model-I extracts and captures features that are used for face recognition. The reason of using a pre-trained model for face recognition is that age estimation from face images and face recognition tasks rely on features extracted from face images. The VGG-Face [99] model is used as the pre-trained CNN model-I. The VGG-Face model has recently been proven to be one of the state-of-the-art in face recognition. The effectiveness of the VGG-Face model comes from the very deep architecture of the model that comprises a big number of convolutional layers. In addition, the VGG-Face model is trained on a very large database consisting of millions of unconstrained images.

As shown in Figure 3, the training data is fed to the VGG-Face model until the values of the first fully connected layer of the VGG-Face model for each example in the training data are calculated. The calculated values of the first fully connected layer represent the facial features that will be used as training data of a new network that has the age labels. This new network consists of four layers. The fourth layer contains the labels. The sizes of the first three layers are 4096, 5000, and 5000 nodes, respectively.

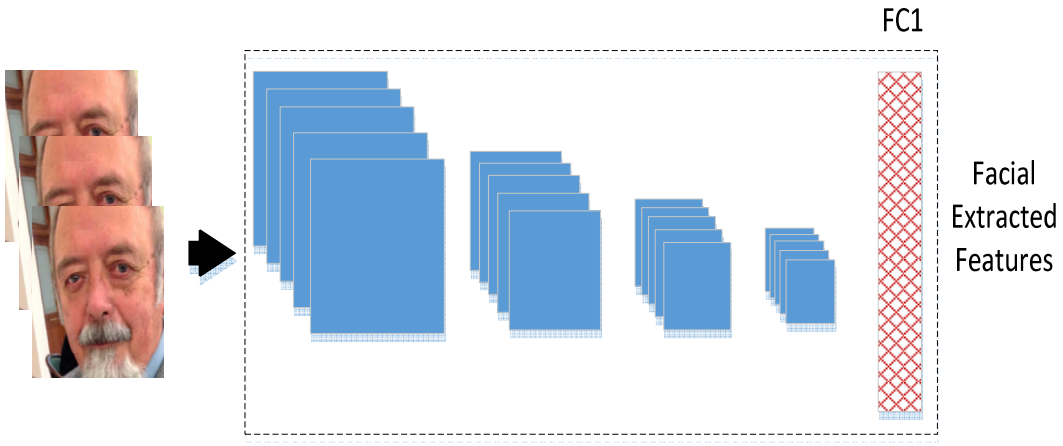

Figure 3. Extraction facial features by using a pre-trained model for face recognition.

3.2.1.2 Pre-trained CNN Model-II

This model aims to find new different features from the face images. Face images are divided into several small regions called superpixels where each superpixel represents a group of pixels as shown in Figure 4. It is assumed that each superpixel is homogeneous in terms of depth, color, and texture meaning all the pixels in a superpixel have near identical depth, color, and texture values. In this model, our main focus is to find information from the superpixels and a relationship between a superpixel and its neighboring superpixels based on their depth. It is assumed that the depth contains age-related information and the depth of the centroid pixel of a superpixel is used to represent the depth of the superpixel. Moreover, variations between the superpixels assist a classifier to find patterns of similarities each age label. This pattern is formed by finding superpixels that have similar depth values.

Three types of similarity measures [107] are used between the superpixel and its neighboring superpixels: the color difference, color histogram, and texture disparity in terms of LBP. In this work, a trained model [108] that finds the features depending on the depth

information is used. The pre-trained model [108] is run on each image example and extracted the last convolutional layer output of this model as new feature set.

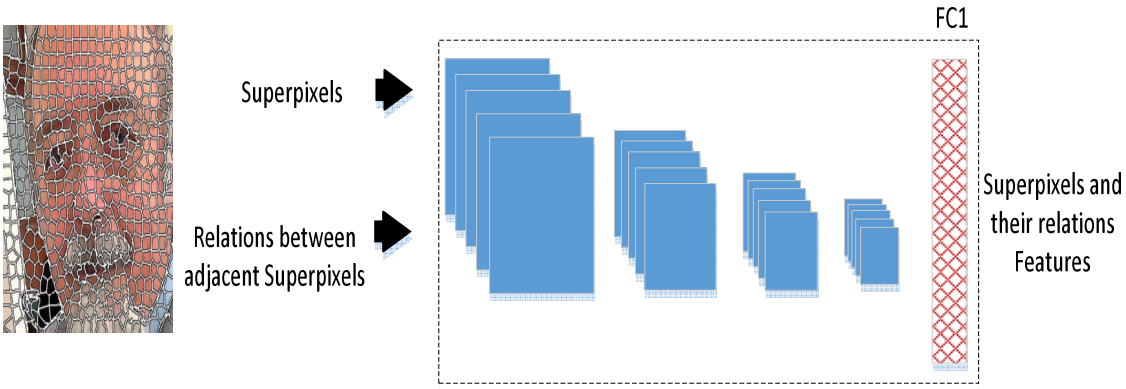

Figure 4. Extraction of superpixels and their relations by using a pre-trained model for depth.

3.2.1.3 Jointly fine-tuned CNN model

The new feature sets extracted from the previous trained models are trained by two new supervised deep neural networks as shown in Figure 5. The first supervised NN, called CNN-F, uses the features extracted by pre-trained CNN Model-I. The second supervised NN, called CNN-S, uses the features extracted by pre-trained CNN Model-II. CNN-F and CNN-S have two hidden layers and one output layer. The third NN, called CNN-FS, jointly fine-tunes CNN-F and CNN-S. CNN-FS is composed of one input layer, one hidden layer, and one output layer. There are 8 labels in the output layer. The input of CNN-FS is the element-wise summation of the last hidden layers of CNN-F and CNN-S. The last hidden layer of CNN-S has 147968 features. Because of the large number of features, dimensionality reduction is used to reduce 25,088 features to 512 features.

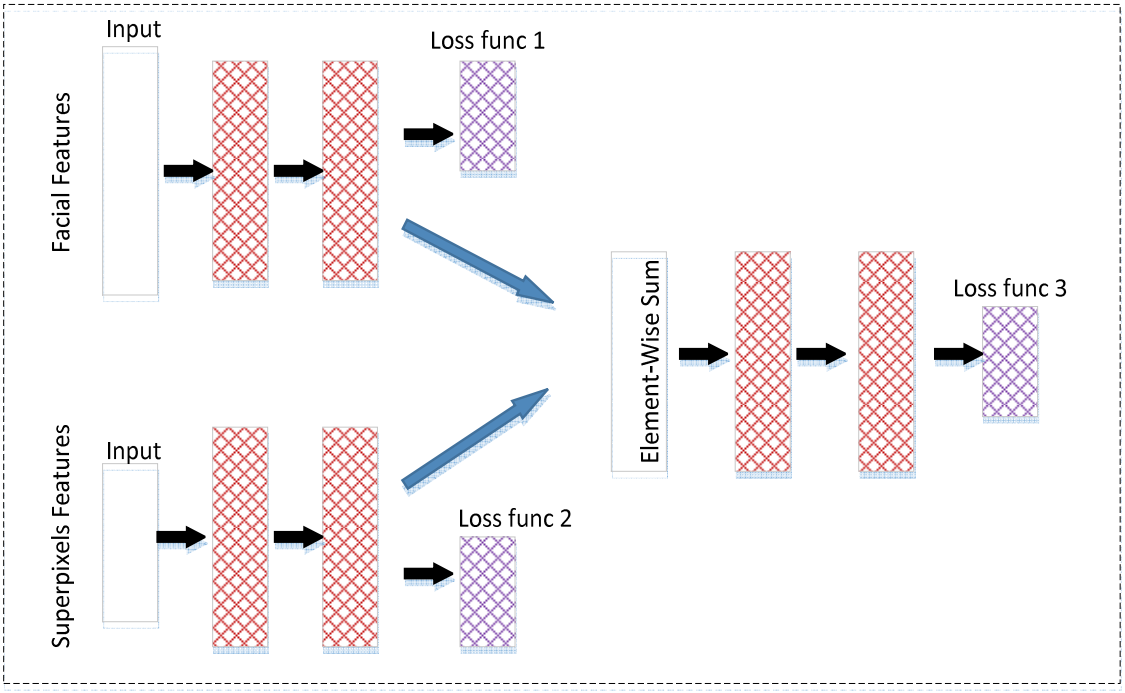

Figure 5. Jointly fine-tuning of two DNNs. CNN-FS uses facial and superpixels related features, the outputs of the last hidden layers of CNN-F and CNN-S, which are summed by element-wise.

The learning process of the three networks is explained as follows.

- Three loss functions are used to train the three networks. All three loss functions are the softmax cross entropy function as in Equation (1).

$$L_i = -\sum_{j=1}^{c} y_j \log(\bar{y}_{i,j}) \quad (1)$$

L_i is the loss function of network i, y_j is the j^{th} value of the label, and $\bar{y}_{i,j}$ is the j^{th} output value of network i.

- CNN-F and CNN-S are trained using their corresponding first batch of features. Then, the output of the last hidden layer of both networks is element-wise summed to form the input of the CNN-FS as in Equation (2).

$$x_{3,j} = Relu(l_{1,j}) + Relu(l_{2,j}) \quad (2)$$

$x_{3,j}$ is the input of the CNN-FS, $l_{1,j}$ and $l_{2,j}$ are the outputs of the last hidden layer of the CNN-F and CNN-S, and *Relu* is the rectified activation function.

- Feedforwarding, error calculation, and backpropagation are performed on the CNN-FS.
- The steps from 1 to 3 are repeated for the rest of the training batches.

After the training is completed, the softmax output of the CNN-FS, $\bar{s}$, is obtained as the final decision as in Equation (3).

$$\bar{s} = \arg max_j \, \bar{y}_{3,j} \tag{3}$$

3.2.2 A New proposed cost function for jointly fine-tuning two different DNNs for age classification

In this section, the proposed new cost function for fine-tuning two DNNs jointly to improve the age and gender classification is explained. The proposed method consists of two DNNs which have two feature sets. These feature sets are extracted for the same data input and each set represents the same input in different way. The first network, DNN1, is trained using the first feature set. The cross-entropy is used as the loss function. The Softmax function is used at the output layer. While the second network, DNN2, is trained on the second feature set. The sigmoid function is used to calculate the output layer probabilities, and the mean squared error loss function is used to calculate the DNN2 error. Then both networks are fine-tuned using the proposed cost function.

3.2.2.1 Architecture of jointly fine-tuning two DNNs and the proposed loss function

In this paper, the proposed integration approach is based on fine-tuning two DNN networks. The first network (DNN1) is a DNN with the first extracted feature set of the input sample, while the input features for the second DNN (DNN2) is the second extracted feature set from the input sample. It is illustrated in Figure 6. Both networks are trained simultaneously and their output functions interact with each other. DNN2 is trained with a sigmoid output function (σ) and the loss mean squared error function (L_{DNN2}) as given by Equations (4) and (5).

$$\bar{y} = \sigma = \frac{1}{1+e^{-z}} \tag{4}$$

$$L_{DNN2} = \frac{1}{n} \times \sum_{j=1}^{n} (y - \bar{y})^2 \tag{5}$$

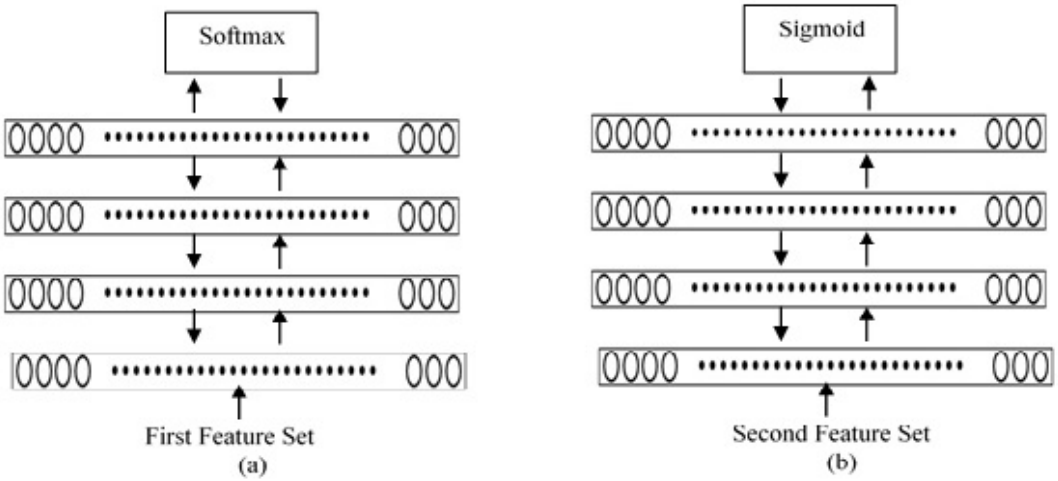

Figure 6. DNN1 and DNN2 architectures. (a) DNN1 with the first extracted feature set as input and with softmax as output layer. (b) DNN2 with Second extracted feature set as input, and with sigmoid as output layer.

z is the input vector for the output layer, n is the number of labels, y is the output vector values of the true label, and $\bar{y}$ is the output vector values of the sigmoid function.

The derivative (*d*) of the loss function is defined in Equation (6), where ⊙ represents the element-wise product.

$$d(L_{DNN2}) = -(y - \bar{y}) \odot \left(\bar{y} \odot (1 - y)\right) \quad (6)$$

DNN1 is trained with the softmax output function and its cross-entropy error function is given by Equations (7) and (8).

$$\bar{y} = Softmax = \frac{e^{z_j}}{\sum e^{z_j}} \quad (7)$$

$$L_{DNN1} = -\sum_j y_j \log \bar{y}_j \quad (8)$$

y_j is the output vector of values of the true label and $\bar{y}_j$ is the output vector values of the softmax function. The derivative (*d*) of the loss function is defined in Equation (9).

$$d(L_{DNN1}) = -(y - \bar{y}) \quad (9)$$

DNN2 is trained and jointly fine-tuned with the DNN1 as shown in Figure 7. The jointly fine-tuned network is trained with the loss function defined in Equation (10).

$$L_{joint} = L_{DNN1} + L_{DNN2} \quad (10)$$

The Softmax and the Sigmoid are the two parts of our proposed joint fine-tuned loss function. The error on the output layer for the jointly fine-tuned network can be calculated by summing the derivatives of both loss function errors of DNN1 and DNN2 as in Equation (11)

$$d\left(L_{Joint}\right) = \left(-(y - \bar{y}) \odot \left(\bar{y} \odot (1 - y)\right)\right) + \left(-(y - \bar{y})\right) \quad (11)$$

Finally, classifying the age and gender of any speaker for any utterance (*S*) is considered by computing the Softmax output values vector, ($\overline{y_J}$), of the jointly fine-tuned network as in Equation (12).

$$S = \arg max_j \bar{y}_j \quad (12)$$

Two loss functions are used jointly to fine-tune the newly proposed method for speaker age and gender classification by using two different feature sets for the same database. The generated error from the first feature set is different than the second set during training. It means that the effect of fine-tuning is different on both DNNs. By correlating the generated error of two related feature sets on the same epoch and on the same batch simultaneously, it is aimed to compute more accurate error value that helps the jointly fine-tuning to reflect an accurate update on the weights and biases of the network. Moreover, the proposed jointly fine-tuned method is based on a new cost function that is derived from two different cost functions, the Softmax and Sigmoid. The Softmax function models the joint distribution over the output variables, which means increasing the value for some outputs leads to the probability of other outputs being decreased. The Sigmoid function models the marginal distributions over the outputs so that increasing or decreasing one of the output values will not affect the other outputs. In this work, the different nature of each function is merged by adding the error generated from the Sigmoid to the Softmax function in order to identify different error sources.

Over-fitting is a problem in machine learning and it can lead the network parameters to over-fit the data from the training samples, leading to failure in classification for the test samples. Different techniques are used to reduce the effect of overfitting such as dropout and weight decay [109] in literature. The proposed loss function helps to minimize the effect of over-fitting by jointly fine-tuning the error of DNN1 and DNN2.

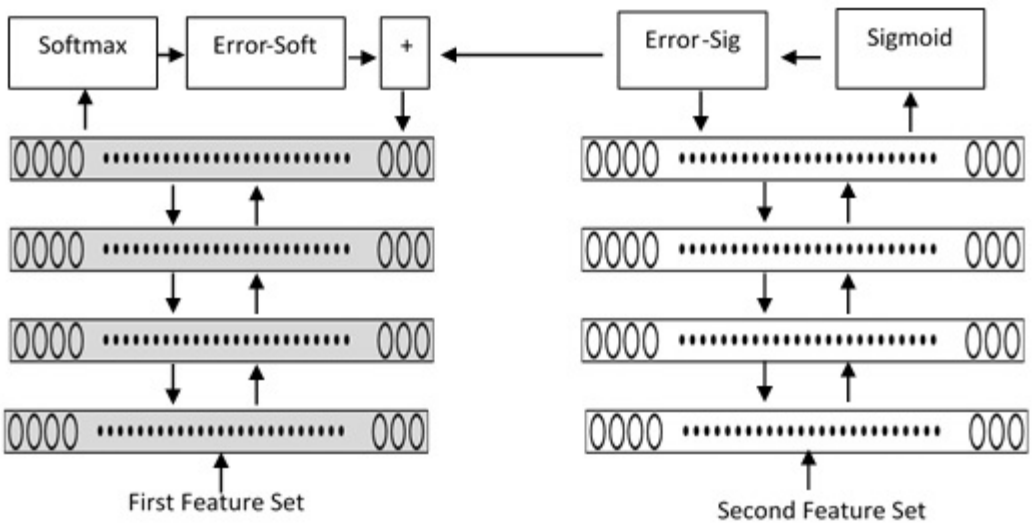

Figure 7. Architecture of the proposed joint fine-tuned DNN1 and DNN2 with the proposed cost function.

3.2.2.2 Score level fusion

Let *n* be the number of the labels of the output layer and the output posterior probabilities of the DNN1 and DNN2 are out1 and out2, respectively. Then, the fused vector *S* for a given utterance *j* can be written by using Equation (13).

$$S_j = \beta \; x \; out_1 + (1 - \beta) \; x \; out_2 \qquad (13)$$

The final scoring, (S_j), is considered to be the index of the maximum value of the system fusion output vector. β is the controlling parameter used for fusing the output results of the two networks. Their values are set based on the accomplishment of each network. After conducting extensive experiments, β is set to be 0.8.

3.3 DNNs and $l_{2,1}$ Norm Regularization for Robust Age Features Selection

Recently, DNNs and $l_{2,1}$ norm [110-112] have been proved to be powerful tools for the feature selection problem. Motivated by these recent advances, a regularization framework is used for selecting features for age classification from facial images, by combining the powerful of the DNN and $l_{2,1}$ norm for extraction the features.

We present the details of our DNN and $l_{2,1}$ model for selecting robust features for age classification. Given data: $[X = x_1, x_{2,\dots}, x_m] \in \mathrm{R}^{\mathrm{fxm}}$, let $[Y = y_1, y_{2,\dots}, y_m] \in \mathrm{R}^{\mathrm{mxc}}$, where m is the number of training samples, f is the number of feature dimension, and c is the number of classes. Our target is to learn a projection matrix $W \in R^{\mathrm{fxc}}$ to select robust features into the common space defined by class labels using $l_{2,1}$ regularized and DNN with sigmoid output layer and mean squared error (MSE) as cost function. Using MSE, sigmoid as output layer function and $l_{2,1}$ regularization, the following minimization objective function is introduced as in Equation (14):

$$E(y,x) = \left\{\frac{1}{2m}(sigm(X^T W) - Y)^2 + \lambda||W||_{2,1}\right\} \quad (14)$$

where $\|.\|_{2,1}$ is an $l_{2,1}$ norm, and $sigm(z) = \left(\frac{1}{(1+e^{-z})}\right)$.

To find the projection matrix W using DNN, Figure 8 shows our DNN model for age feature selection. As it can be seen the model is composed of 2 hidden layers with f nodes in each layer which is equal to the number of nodes in the input layer.

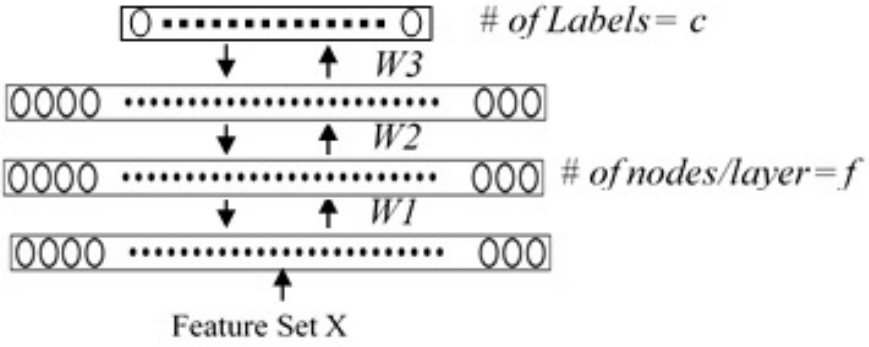

Figure 8. DNN architecture for finding the projection matrix W.

The number of nodes in the output layer is set to the number of labels (c). After training the model, W3 is taken as our projection matrix W.

3.3.1 Learning

In the learning process, it is aimed to minimize the proposed cost function with respect to the projection matrix W as in Equation (15).

$$min_W^{E(y,x)} = \{(sigm(X^T W) - Y)^2 + \lambda ||W||_{2,1}\} \quad (15)$$

With the presence of the $l_{2,1}$ norm the objective function in (15) is not easy to minimize. However, in [111-112], they proposed to solve the minimization of the $l_{2,1}$ norm based on the half-quadratic minimization. As well as, one should know that the minimizer function of $l_{2,1}$ norm is unpredictable near the origin. Therefore, according to $l_{2,1}$ norm analysis in [111], a $\emptyset(x) = \sqrt{\epsilon + x^2}$ can be defined to solve this problem, where ϵ is chosen to be a decreased value to ensure that the function in (15) with $l_{2,1}$ norm is converged. And $\emptyset$ should satisfy all the conditions in (16).

$$x \to \emptyset(x)\ is\ convex\ on\ R,$$

$$x \to \emptyset\left(\sqrt{x}\right)\ is\ concave\ on\ R_+,$$

$$\emptyset(x) = \emptyset(-x), \forall x \in R, \tag{16}$$

$$\emptyset(x)\ is\ C^1\ on\ R,$$

$$\emptyset''(0^+) > 0, \lim_{x\to\infty} \emptyset(x)/x^2 = 0$$

Lemma 1. Let $\emptyset(x)$ be a function satisfying all condition in (16), there exist a conjugate function $\varphi(.)$ as in Equation (17) such that

$$\emptyset\left(\left|\left|w^i\right|\right|_2 = \inf_{p\in R}\left\{p\left|\left|w^i\right|\right|_2^2 + \varphi(p)\right\}\right. \tag{17}$$

where p is determined by the minimizer function $\delta(.)$ with respect to $\emptyset(.)$.

Based on $\emptyset(x)$, $\lambda||W||_{2,1}$ is replaced with $\lambda\sum_i^f \sqrt{\epsilon + ||w^i||_2^2}$, then the equation in (15) is formulated as in Equation (18).

$$\min_W^{E(y,x)} = \left\{\frac{1}{2m}(sigm(X^TW) - Y)^2 + \lambda\sum_i^f \sqrt{\epsilon + ||w^i||_2^2}\right\} \tag{18}$$

According to Lemma 1, the function of $\lambda\sum_i^f \sqrt{\epsilon + ||w^i||_2^2}$ can be reformulated as in Equation (19):

$$\lambda Tr(W^TQW) \tag{19}$$

where q =$\delta\left(||w^i||_2\right) \in R^f$ is an auxiliary vector, and Q = $diag$(q). The operator $diag(.)$ puts a vector q on the main diagonal of Q. q is computed using the optimizer function as in Equation (20).

$$q_i = \frac{1}{\sqrt{||w^i||_2^2+\epsilon}} \tag{20}$$

According to (19), our minimization function can be written as in Equation (21),

$$\min_{W} E(y,x) = \frac{d}{dW}\left\{\frac{1}{2m}(sigm(X^T W) - Y)^2 + \lambda Tr(W^T Q W)\right\} \quad (21)$$

The analytic minimization solution of (21) with respect to W is given by Equation (22),

$$\min_{W} E(y,x) = (sigm(X^T W) - Y)\left(sigm(X^T W)\left(1 - \frac{1}{sigm(X^T W)}\right)\right) + \lambda Q W \quad (22)$$

CHAPTER 4: EXPERIMENTAL SETTINGS AND DNNs CONFIGURATIONS

Several experiments have been conducted to evaluate the proposed methods and techniques. A publicly available database of age from unconstrained facial images is used to evaluate the effectiveness of the proposed methods compare with the state-of-the-art. The training and testing of our experiments have been developed using MatConvNet and DeepLearmMaster Toolboxes with our own modifications. The computation time for training DNNs is very high. INVIDIA TITAN X GPU with 3072 cores and 12 GB of video memory is used to accelerate the training time.

4.1 Benchmark

The Adience benchmark is used in this work. The Adience, contains 26K face images of 2284 subjects who are divided into 8 age groups called labels. Table 3 shows the Adience labels and the number of images per label. Standard five-fold, subject-exclusive cross-validation protocol is applied for dividing the database into a train and test groups. The same settings were used in [64]. The Adience is a challenging database since it consists of unfiltered face images, which were uploaded to the Flicker website using smart phones. The images are not filtered with any manual filtering techniques. Images in the database reflect real-world conditions of uncontrolled environments such as significant variations in pose, expression, lighting, image quality and resolution.

Table 3. The Adience benchmark.

	Labels (in years)								Total
Gender	0-2	4-6	8-13	15-20	25-32	38-43	48-53	60-	**# of im.**
F	682	1234	1360	919	2589	1056	433	427	9411
M	745	928	934	734	2308	1294	392	442	8192

The Adience is not designed for face recognition task so that the number of images per subject is not balanced. Around 80 percent of the subjects in the database have only one image, while the rests have around 100 to 400 images. When the number of images per subject is small for a label while it is bigger for other labels, the classifier will be biased for the labels with more images.

4.2 Settings for Jointly Fine-Tuning DNNs Based on Amplified Feature Sets

Three supervised neural networks, CNN-F, CNN-S, and CNN-FS are trained separately. The training settings for the CNN-F and CNN-S are summarized in Table 4. These two networks are jointly fine-tuned from scratch to design the third network CNN-FS. On the top of the last hidden layers of the CNN-F and CNN-S, two more hidden layers are added of the size of 1024 and 256, respectively. Then, one output layer with eight age labels is added.

The learning rate for the CNN-FS is set to 0.01, with dropout rate of 0.8, weight decay of 10-4. The training is stopped when there is no improvement in the validation set results.

Table 4. Network settings for CNN-F and CNN-S.

	Input layer	Hidden layers	Learning rate	Dropout	Weight decay
CNN-F	4096	1024, 1024	0.1	0.7	10^{-3}
CNN-S	512	1024, 1024	0.1	0.7	10^{-3}

4.3 Settings and Configurations for Jointly Fine-Tuning DNNs Using the Proposed Cost Function

The input feature set for DNN1 is the facial information obtained from the pre-trained VGG-Face model for face recognition and the number of nodes in the input layer for DNN1 is 4096. While the input feature set for DNN2 is the depth information which obtained from the superpixels and their relation and the number of nodes in the input layer for DNN2 is 512. the network settings for the DNN1, DNN2, and the joint fine-tuned DNNs is shown in Table 5.

Table 5. Proposed network architectures and settings.

	DNN1	DNN2	Joint Fine-Tuned DNNs
No. of Hidden layers	2	2	2
No. of Nodes/hidden layer	1024	1024	512
Learning rate	0.1	0.1	Start at 0.1 then decreased by 0.2 every 3 epochs
Dropout	0.7	0.7	0.5
Weight Decay	10^{-3}	10^{-3}	10^{-4}
No. of Epochs	15	15	20
Cost Function	Softmax	Sigmoid	L_{joint}
Input features	Facial	Superpixel	Facial+Superpixel
Activation Function	Rectified	Rectified	Rectified

4.4 Robust Feature Selection Training Settings

The proposed work is applied on the facial features that were extracted from the Adience database and obtained the projection matrix of features weights. The projection

matrix is mapped on the feature set to select the robust features. After that, the facial features and the new robust features are concatenated and trained for age classification. The settings for finding the projection matrix, training the concatenated features, and for training the robust feature set are shown in Table 6.

Table 6. Training settings for different DNN networks.

Network	# of hidden layers	# of input features	# of nodes in each layer	Dropout rate	Weight decay	Learning Rate
DNN for finding the projection matrix	2	4096	4096-4096	0.7	10^{-4}	0.01
DNN for training the concatenated features	2	4104	512-512	0.7	10^{-4}	0.01
NN for training the robust features alone	1	8	50	0.5	10^{-4}	0.01

The parameter λ can be calibrated automatically or manually. In our case, the best value is found to be 0.1 in all experiments.

CHAPTER 5: DISCUSSION AND EXPERIMENTAL RESULTS

5.1 Pre-trained CNN Models Feature Sets Classification Results

The accuracy results for using the pre-trained models for age estimation are shown in Table 7. After fine-tuning each network from their original task to age classification task, GoogLeNet, ResNet-50, VGG-VD-16, VGG-VD-19, and FNC-8s achieved 45.07%, 42.46%, 45.01%, 45.99%, and 43.87% of accuracy, respectively. Fine-tuning from different classification tasks to age classification provides reasonable accuracies compared to the state-of-art results. The highest accuracy (57.45%) is achieved by using VGG-Face model. It is noticed that the employment and retraining of a very deep and well-trained CNN for face recognition improves the performance of age estimation better than the other pre-trained models. Confusion matrix for utilizing VGG-Face for age estimation is presented in Table 8. These results support the fact that both the number of the training images and subjects of the used database and the pre-training task of the CNN determine the network ability to achieve good results for age classification from facial images.

Table 7. Overall accuracies of different CNN architectures (%).

Label	GoogLeNet	ResNet-50	VGG-VD-16	VGG-VD-19	FNC-8s	VGG-Face
0-2	86.75	90.89	84.27	83.44	79.92	88.41
4-6	27.89	15.79	42.28	43.68	30.35	60.18
8-13	21.47	0.29	33.82	30.59	29.71	39.12
15-20	14.10	0.00	14.98	11.89	18.50	43.61
25-32	76.61	97.82	58.90	62.03	66.29	67.14
38-43	12.03	0.00	24.06	27.42	17.36	43.79

48-53	7.05	0.00	12.45	10.79	4.98	14.52
60-	34.63	0.00	33.46	35.02	43.97	57.2
Overall Acc.	**45.07**	**42.46**	**45.01**	**45.99**	**43.87**	**57.45**

Table 8. Confusion matrix for the fine-tuning VGG-Face for age estimation (%).

Predicted / *Actual*	**0-2**	**4-6**	**8-13**	**15-20**	**25-32**	**38-43**	**48-53**	**60-**
0-2	**88.41**	10.56	0.21	0.00	0.83	0.00	0.00	0.00
4-6	24.04	**60.18**	12.46	2.46	0.88	0.00	0.00	0.00
8-13	0.88	11.18	**39.12**	36.47	11.18	0.59	0.59	0.00
15-20	0.88	0.44	9.69	**43.61**	43.17	2.20	0.00	0.00
25-32	0.00	0.09	1.89	9.66	**67.14**	19.13	1.61	0.47
38-43	0.20	0.20	0.79	2.96	37.67	**43.79**	11.83	2.56
48-53	0.00	0.00	0.83	0.00	7.88	58.51	**14.52**	18.26
60-	0.00	0.00	0.00	1.56	1.56	10.12	29.57	**57.20**

From Table 7, it is observed that networks previously trained for image classification except ResNet-50 perform better than the FNC-8s which was trained for semantic segmentation. It can be stated that CNNs trained for image classification can contain more age-related features compared to CNNs trained for image semantic segmentation. Although ResNet-50 was trained for images classification, its performance, in terms of overall accuracy, is slightly worse than the other CNNs trained for the same task. Moreover, the accuracy results of ResNet-50 for 8-13, 38-43, 48-53, 60- age groups are zero. This might be due to the deep architecture of the ResNet-50 network.

Building an efficient large database containing millions of face images for age classification is a difficult task due to the fact that this requires an access to participants' private information and requires IRB approval to do so. Furthermore, the collected images need to be manually labeled. Therefore, databases that are designed for age classification from dynamic and real environments such as social sites are limited on their size. They are

not comparable in size with other databases which have been designed for object recognition or face recognition such as ImageNet [113] and Pascal [114] databases.

Overfitting is one of the biggest challenges of machine learning, especially when using small databases. It becomes an even more common and significant issue in DNNs, where the networks often have a large number of layers containing thousands of neurons. Hence, the number of connections in these networks is astronomical, reaching to the millions. As a result, a compact architecture network should be designed to trade-off between overfitting and network complexity. If the network is not complex enough, it may not be powerful to capture the necessary information to gain more accurate result. In this work, large databases that are originally formed and used for other recognition tasks such as face recognition in order to estimate the age information from face images are used in advantage.

The following results are obtained after different experiments are conducted to evaluate the effectiveness of the dimensionally reduced features which are extracted based on the pre-trained models for different tasks. The accuracy results of different CNN models and their combinations are given in Table 9 accompanied by the DNN specifications used for each experiment. Table 10 provides the confusion matrix for the model that achieved the highest accuracies. Since the optimal configuration settings for CNNs are problem dependent, different settings have been tested to reach the best results for age estimation from facial images. Four fully connected layers for all models are tested to be the best for achieving the highest accuracy (Table 9). The number of nodes in each layer is different between the layers and between different trained models. For some models, the number of nodes in the first fully connected layer is greater than the number of nodes in the rest of the

fully connected layers, while for other models, the number of nodes in the input fully connected layer is smaller than the rest of the fully connected. For the FNC-8s model, the number of nodes in each fully connected layer is the same. Optimum dropout and weight decay values are chosen after extensive experiments. For the GoogLeNet and FNC-8s models, the dropout rate value is chosen as 0.5, while it is set to 0.8 for the other models. For the weight decay value, different values are tested, ranged from 10^{-5} to 10^{-2}. Most of the models performed well at 10^{-3} while for the GoogLeNet achieved the best performance at 10^{-4}.

In general, it is observed from the Tables 7 and 9 that the dimensionality reduction improves the classification performance of all CNNs models for the age classification task. It is also observed that the best results are achieved with the VGG-Face CNN model. Another observation is the significant improvement by the FNC-8s with dimensionality reduction. FNC-8s achieved the highest accuracy among the other networks except the VGG-Face. Moreover, the performances get better when all features of different networks are combined together. For example, combining the features of the FNC-8s with the features of the VGG-VD-19 achieved better results than their individual use. The best results are achieved when all networks are combined with dimensionality reduction.

Table 9. Overall accuracies of different CNN models with dimensionality reduction.

Trained Model	Accuracy (%)	# of fully connected layers	# of nodes in each layer	Dropout rate	Weight decay
VGG-Face	60.60	4	512-1024-1024	0.8	10^{-3}
GoogLeNet	46.43	4	512-1024-1024	0.5	10^{-4}
ResNet-50	45.69	4	512-1024-1024	0.5	10^{-4}
VGG-VD-16	47.13	4	512-1024-1024	0.8	10^{-3}
VGG-VD-19	47.49	4	512-1024-1024	0.8	10^{-3}
FNC-8s	48.95	4	512-512-512	0.5	10^{-3}
VGG-Face + FNC-8s	61.39	4	1024-512-512	0.8	10^{-3}

VGG-VD-19+ FNC-8s	51.88	4	1024-512-512	0.8	10^{-3}
Combined Models (VGG-Face+ GoogLeNet+ ResNet-50+ VGG-VD-16+ VGG-VD-19+ FNC-8s)	62.26	4	3072-1024-1024	0.8	10^{-3}

The network settings for each experiment are indicated.

Table 10. Confusion matrix for the highest accuracy combined models with dimensionality reduction (%).

Predicted / *Actual*	**0-2**	**4-6**	**8-13**	**15-20**	**25-32**	**38-43**	**48-53**	**60-**
0-2	**86.54**	12.84	0.41	0.00	0.21	0.00	0.00	0.00
4-6	25.09	**62.28**	10.18	1.05	0.70	0.18	0.00	0.53
8-13	0.59	7.06	**45.88**	22.65	20.88	2.35	0.00	0.59
15-20	0.00	0.44	8.37	**31.28**	56.39	3.52	0.00	0.00
25-32	0.19	0.19	2.37	3.13	**77.37**	16.29	0.28	0.19
38-43	0.00	0.20	0.59	2.76	35.70	**47.14**	3.55	10.06
48-53	0.00	0.00	0.83	0.00	9.54	55.60	**11.62**	22.41
60-	0.00	0.00	0.00	0.00	2.33	10.89	7.00	**79.77**

Figure 9. shows some of the challenging images from the Adience database. These images are classified correctly by using the proposed work although they consist of image formation distortions such as motion blur, low-resolution, pose, and facial expressions.

Figure 9. Some of the challenging images classified correctly by this work.

5.2 The Proposed Amplified Feature Sets Results and Discussion

In this section, the performance of the Joint Fine-Tuning DNNS Based on Amplified Feature Sets is evaluated, several experiments have been carried out. The evaluation of the CNN-F and CNN-S and CNN-FS networks is investigated separately.

Table 11 presents the overall accuracies of these networks. Tables 12,13, and 14 show the confusion matrices for the CNN-F, CNN-S, and CNN-FS, respectively.

In Table 11, it is noticed that the overall accuracy of the jointly tuned network, the CNN-FS, outperforms the accuracy of the CNN-F and CNN-S about 7%. Although the CNN-F and CNN-S networks performs well, the performance of the CNN-F is slightly better than that of the CNN-S. This might be due to two reasons: 1) The CNN-F was trained on features that were extracted from a model that was pre-trained on a database which has millions of images while the CNN-S was trained on features extracted from a model that was pre-trained on a database that has a few thousands of images. 2) The facial features may have more age-related information than the features extracted from the superpixels and their relation with the adjacent superpixels. It is also noticed that the performance of the CNN-F network was better than the CNN-FS joint network for some classes. This happened when the classification accuracy of the CNN-S was less than that of the CNN-F. The element-wise summation layer of the jointly fine-tuned network works as an amplifier. Whenever the CNN-F and CNN-S had good performance for a class together, the performance of the CNN-FS network got better as it happened for class 8 (60-) and class 3 (8-13). On the other hand, when the performances of the CNN-F or CNN-S networks were apart from each other, the CNN-FS network was affected by the network with a lesser performance. For example, class 1 (0-2), class 2 (4-6), class 4 (15-20), class 5 (25-32), class 6 (38-43), and class 7 (48-53).

Table 11. Overall classification accuracies of the CNN-F, CNN-S, CNN-FS (%).

	CNN-F	**CNN-S**	**CNN-FS**
0-2	88.40	62.33	87.99
4-6	60.17	58.82	60.70
8-13	39.12	34.17	47.65

15-20	43.61	15.01	39.20
25-32	67.14	80.77	80.49
38-43	43.79	11.04	47.53
48-53	14.52	10.88	12.86
60-	57.20	53.30	79.37
Overall Accuracy %	57.45	53.62	63.78
1-off Accuracy %	94.32	81.40	93.70

Table 12. Confusion matrix for the CNN-F (%).

Predicted / *Actual*	**0-2**	**4-6**	**8-13**	**15-20**	**25-32**	**38-43**	**48-53**	**60-**
0-2	**88.41**	10.56	0.21	0.00	0.83	0.00	0.00	0.00
4-6	24.04	**60.18**	12.46	2.46	0.88	0.00	0.00	0.00
8-13	0.88	11.18	**39.12**	36.47	11.18	0.59	0.59	0.00
15-20	0.88	0.44	9.69	**43.61**	43.17	2.20	0.00	0.00
25-32	0.00	0.09	1.89	9.66	**67.14**	19.13	1.61	0.47
38-43	0.20	0.20	0.79	2.96	37.67	**43.79**	11.83	2.56
48-53	0.00	0.00	0.83	0.00	7.88	58.51	**14.52**	18.26
60-	0.00	0.00	0.00	1.56	1.56	10.12	29.57	**57.20**

Table 13. Confusion matrix for the CNN-S (%).

Predicted / *Actual*	**0-2**	**4-6**	**8-13**	**15-20**	**25-32**	**38-43**	**48-53**	**60-**
0-2	**86.34**	12.84	0.21	0.00	0.41	0.00	0.00	0.21
4-6	24.30	**58.82**	8.25	3.86	3.19	0.35	0.00	1.23
8-13	1.76	12.06	**34.18**	10.00	37.00	2.35	0.00	2.65
15-20	0.00	4.41	15.42	**15.01**	57.23	4.41	1.32	2.20
25-32	0.57	1.33	4.17	5.59	**80.78**	4.07	1.23	2.27
38-43	0.20	2.56	2.96	3.75	62.13	**11.05**	6.90	10.45
48-53	0.00	3.32	2.49	5.81	47.62	14.94	**10.88**	14.94
60-	0.39	1.56	2.72	5.45	24.90	7.78	3.89	**53.31**

Table 14. Confusion matrix for the CNN-FS (%).

Predicted / *Actual*	**0-2**	**4-6**	**8-13**	**15-20**	**25-32**	**38-43**	**48-53**	**60-**
0-2	**87.99**	11.59	0.21	0.00	0.21	0.00	0.00	0.00
4-6	21.40	**60.70**	16.49	0.70	0.70	0.00	0.00	0.00
8-13	0.59	7.06	**47.65**	23.53	20.00	1.18	0.00	0.00
15-20	0.00	0.00	10.13	**39.21**	49.34	1.32	0.00	0.00
25-32	0.00	0.09	1.70	5.21	**80.49**	11.93	0.47	0.09
38-43	0.00	0.20	1.38	2.17	38.86	**47.53**	2.37	7.50
48-53	0.00	0.00	0.00	0.83	11.20	52.70	**12.86**	22.41
60-	0.00	0.00	0.00	1.56	1.17	10.51	7.39	**79.38**

Figure 10 top row shows a set of images that are incorrectly classified by the proposed networks. Figure 10 bottom row shows a set of images that are classified correctly by the proposed networks. It can be observed that for a certain extent the proposed methods are capable to estimate the correct age of a person from face images regardless of the challenging nature of images such as poor resolution and pose. Our proposed networks failed to classify images correctly in case of extreme blur, low-resolution, pose, and alignment in images.

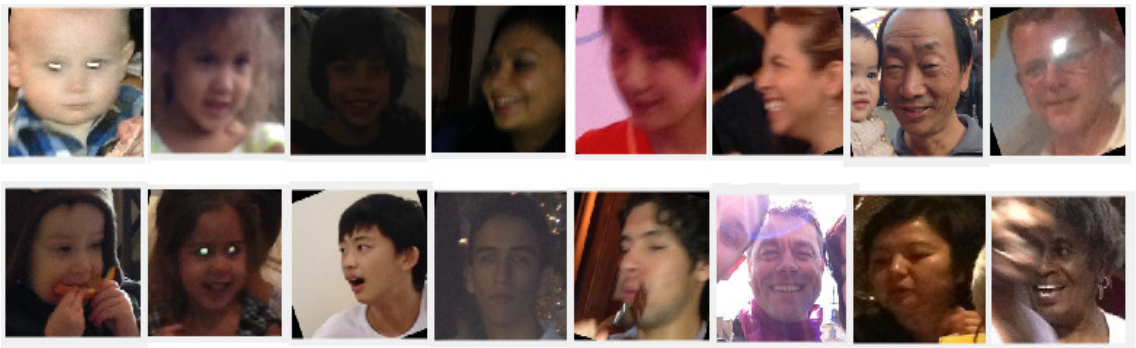

Figure 10. Challenging images in the Adience database. Images in the top row were classified correctly by the proposed networks. Images in the bottom row were classified incorrectly by the proposed networks.

5.3 The Classification Results of the Proposed Cost Function

To demonstrate the validation of the performance of the proposed networks and the cost function furthermore, several experiments have been carried out for age classification from face image database. The evaluation of the DNN1 with the extracted facial features as input, DNN2 with depth from superpixels features as input, and the Joint Fine-Tuned with the proposed cost function network, is investigated separately. Table 15 presents the overall accuracies of these networks. As well as Tables 16, 17, and 18 show the confusion matrices for the DNN1, DNN2, and the Joint Fine-Tuned network, respectively.

Table 15 shows the exact group and the 1-off (when the person belongs to his/her exact group or the group immediately before or after his/her exact group) classification accuracy for all age groups. It is noticed that the overall accuracy of the jointly tuned network using the new proposed cost function, outperforms the accuracy of the DNN1 and DNN2 in both cases the exact and the 1-off accuracy.

Table 15. Overall classification accuracies for the proposed cost function on Adience database (%).

	0-2	**4-6**	**8-13**	**15-20**	**25-32**	**38-43**	**48-53**	**60-**	**Accuracy%**	**1-off Acc**
DNN1	88.41	60.18	39.12	43.61	67.14	43.79	14.52	57.20	**57.45**	**94.32**
DNN2	86.34	58.82	34.18	15.01	80.78	11.05	10.88	53.31	**53.62**	**81.40**
Joint Fine-Tuned	**85.92**	**62.28**	**45.29**	**36.12**	**76.70**	**44.97**	**14.52**	**84.44**	**62.37**	**94.46**

Table 16. Confusion matrix for DNN1 with facial features as input (%).

Predicted / *Actual*	**0-2**	**4-6**	**8-13**	**15-20**	**25-32**	**38-43**	**48-53**	**60-**
0-2	**88.41**	10.56	0.21	0.00	0.83	0.00	0.00	0.00
4-6	24.04	**60.18**	12.46	2.46	0.88	0.00	0.00	0.00
8-13	0.88	11.18	**39.12**	36.47	11.18	0.59	0.59	0.00
15-20	0.88	0.44	9.69	**43.61**	43.17	2.20	0.00	0.00
25-32	0.00	0.09	1.89	9.66	**67.14**	19.13	1.61	0.47
38-43	0.20	0.20	0.79	2.96	37.67	**43.79**	11.83	2.56
48-53	0.00	0.00	0.83	0.00	7.88	58.51	**14.52**	18.26
60-	0.00	0.00	0.00	1.56	1.56	10.12	29.57	**57.20**

Table 17. Confusion matrix for DNN2 with depth features based on superpixels and their relations as input (%).

Predicted / *Actual*	**0-2**	**4-6**	**8-13**	**15-20**	**25-32**	**38-43**	**48-53**	**60-**
0-2	**86.34**	12.84	0.21	0.00	0.41	0.00	0.00	0.21
4-6	24.30	**58.82**	8.25	3.86	3.19	0.35	0.00	1.23
8-13	1.76	12.06	**34.18**	10.00	37.00	2.35	0.00	2.65
15-20	0.00	4.41	15.42	**15.01**	57.23	4.41	1.32	2.20
25-32	0.57	1.33	4.17	5.59	**80.78**	4.07	1.23	2.27
38-43	0.20	2.56	2.96	3.75	62.13	**11.05**	6.90	10.45
48-53	0.00	3.32	2.49	5.81	47.62	14.94	**10.88**	14.94
60-	0.39	1.56	2.72	5.45	24.90	7.78	3.89	**53.31**

Table 18. Confusion matrix for jointly fine-tuned network with facial and depth features as input, by using the new proposed cost function (%).

Predicted / *Actual*	**0-2**	**4-6**	**8-13**	**15-20**	**25-32**	**38-43**	**48-53**	**60-**
0-2	**85.92**	12.63	0.41	0.00	1.04	0.00	0.00	0.00
4-6	23.16	**62.28**	12.11	1.75	0.53	0.00	0.00	0.18
8-13	0.59	7.65	**45.29**	30.59	14.71	0.88	0.00	0.29
15-20	0.00	0.44	11.45	**36.12**	50.66	1.32	0.00	0.00
25-32	0.00	0.09	1.61	3.13	**76.70**	17.14	0.95	0.38
38-43	0.20	0.00	0.79	0.99	37.67	**44.97**	7.50	7.89
48-53	0.00	0.00	0.00	0.00	9.13	39.83	**14.52**	36.51
60-	0.00	0.00	0.00	1.17	0.78	5.45	8.17	**84.44**

It can be observed that the performance of DNN1 with the facial feature as input is bit better than the performance of DNN2 which takes the depth features which extracted based on the image superpixels and their relation. This might be due to the fact that the facial features has more significant information related to the age of human. As well as, it is noticed that the performance of network which trained only using facial features in some age groups gives better or almost same results likes the Joint Fine-Tuned netwok such as in age groups (15-20), (48-53). This clearly happenes when the trained newtwork with the depth featuers extracted based on the superpixels is giving relatively poor results for the same age groups. However, the results of training both networks jointly using the proposed cost function indicates that a significant improvement in the accuracy for age classification is notable as in the case of classifying the speaker age and gender using the proposed cost function.

5.4 Robust Features Selection Method Classification Results

To show the effectiveness of the proposed robust feature selection method, the method is tested on the Adience database. Table 19, 20, and 21 show the overall accuracies

and the confusion matrices for the facial features which were obtained using a face recognition pre-trained model, the overall accuracies of the robust features extracted using the proposed method, and the overall accuracies for both feature sets when they are concatenated. As it can be seen from the results in Table 19, the classification results using the robust features are comparable to the classification results of the source features where the robust features originated from. The size of the robust features is eight features which is much smaller than the other features, and this indicates the powerfulness of the proposed method to select the most related features for the age from the input image. As well as, the small size of the robust features requires a relatively smaller network for training and this will result in less computational time.

Table 19. Overall classification accuracies for facial robust features on Adience database.

	0-2	**4-6**	**8-13**	**15-20**	**25-32**	**38-43**	**48-53**	**60-**	**Accuracy (%)**	**1-off Acc (%)**
Facial Features	88.41	60.18	39.12	43.61	67.14	43.79	14.52	57.20	**57.45**	**94.32**
Robust Features	82.19	69.12	39.71	12.78	76.42	19.53	6.64	38.52	**53.25**	**81.18**
Concatenated Features	86.96	65.96	45.88	35.24	78.98	41.22	15.35	83.66	**63.22**	**94.38**

Table 20. Confusion matrix for the facial robust features (%).

Predicted / *Actual*	**0-2**	**4-6**	**8-13**	**15-20**	**25-32**	**38-43**	**48-53**	**60-**
0-2	**82.19**	15.73	0.21	0.00	1.86	0.00	0.00	0.00
4-6	13.51	**69.12**	12.98	1.23	2.81	0.18	0.00	0.18
8-13	0.59	5.88	**39.71**	5.59	44.12	3.82	0.00	0.29
15-20	0.44	0.88	15.42	**12.78**	65.64	3.96	0.00	0.88
25-32	0.19	0.38	5.49	1.52	**76.42**	11.65	0.76	3.60
38-43	0.59	0.39	1.97	0.59	62.73	**19.53**	3.55	10.65
48-53	0.83	0.41	3.73	2.49	52.70	23.24	**6.64**	9.96
60-	0.00	0.39	1.17	0.00	29.97	27.24	2.72	**38.52**

Table 21. Confusion matrix for the facial features concatenated with their robust features (%).

Predicted / *Actual*	**0-2**	**4-6**	**8-13**	**15-20**	**25-32**	**38-43**	**48-53**	**60-**
0-2	**86.96**	12.01	0.41	0.00	0.62	0.00	0.00	0.00
4-6	20.70	**65.96**	11.23	1.40	0.70	0.00	0.00	0.00
8-13	0.59	7.06	**45.88**	30.29	15.29	0.59	0.00	0.29
15-20	0.44	0.44	11.45	**35.24**	51.10	1.32	0.00	0.00
25-32	0.00	0.09	1.61	3.98	**78.98**	14.30	0.57	0.47
38-43	0.00	0.00	0.99	1.18	41.62	**41.22**	5.72	9.27
48-53	0.00	0.00	0.83	0.00	8.71	39.00	**15.35**	36.10
60-	0.00	0.00	0.00	1.17	0.78	5.06	9.34	**83.66**

The concatenation of the original features and their selected robust features improves the classification results by a wide margin. From the confusion matrices in Table 20 and Table 21 it is noticed that the misclassification ratio for the two feature sets are not the same for some classes. As an example, class (60-) the misclassification for the facial feature set occurs mainly in class (48-53) while the misclassification for the robust features occurs in class (25-32) and class (38-43). Therefore, the concatenation of the two features helps to get better accuracy for such classes.

5.5 Utilizing the New Cost Function and the Jointly Fine-Tuned Amplified Network Using the Proposed Robust Features

In this section, the proposed robust feature selection method is applied to extract the robust features of the new feature set which is based on the superpixels and their relations. The process of extracting this feature set is explained in section 3.3. The resulted two feature sets after applying the proposed robust selection method namely, the robust-facial features and the robust-superpixels features are fine-tuned using the two jointly fine-tuning methods proposed in sections 3.2.1 and 3.2.2. The robust-facial features were

extracted in section 5.4 using the robust feature selection method, the same method is used to extract the robust-superpixel features.

Table 22 shows the overall accuracies for the superpixels features which were obtained using a depth estimation pre-trained model, the overall accuracies of the robust-superpixel features, and the overall accuracies for the superpixels features and their robust features when they are concatenated. Tables 23 and 24 show the confusion matrix for the robust-superpixels features and the concatenation of the superpixels features with their robust features. As it can be seen from these tables, the classification results using the robust-superpixel features are comparable to the classification results of the superpixels features. And it can be observed that the concatenation of both features improves the classification results.

Table 22. Overall classification accuracies for superpixels robust features on Adience database (%).

	0-2	**4-6**	**8-13**	**15-20**	**25-32**	**38-43**	**48-53**	**60-**	**Overall Accuracy**	**1-off Acc**
Superpixels Features	86.34	58.82	34.18	15.01	80.78	11.05	10.88	53.31	**53.62**	**81.40**
Robust-Superpixels Features	78.14	52.37	36.46	20.28	67.61	17.77	17.38	57.15	**78.89**	**49.95**
Concatenated Features	87.56	56.82	43.57	31.05	78.31	24.81	29.70	60.96	**58.31**	**86.17**

Table 23. Confusion matrix for the superpixels robust features (%).

Predicted / *Actual*	**0-2**	**4-6**	**8-13**	**15-20**	**25-32**	**38-43**	**48-53**	**60-**
0-2	**78.14**	17.35	2.24	0.32	1.41	0	0	0.54
4-6	28.27	**52.37**	9.42	4.64	2.92	1.16	0.64	0.58
8-13	2.31	7.67	**36.46**	8.54	38.12	5.03	0	1.87
15-20	0	4.99	14.08	**20.28**	53.24	3.92	1.85	1.64
25-32	1.88	2.43	6.04	7.93	**67.61**	9.51	1.23	3.37
38-43	1.63	2.66	3.49	6.7	55.44	**17.77**	5.08	7.23
48-53	0	3.96	5.65	3.72	44.29	11.89	**17.38**	13.11
60-	0.07	2.07	3.69	7.81	20.83	5.46	2.92	**57.15**

Table 24. Confusion matrix for the superpixels features concatenated with their robust features (%).

Actual \ *Predicted*	**0-2**	**4-6**	**8-13**	**15-20**	**25-32**	**38-43**	**48-53**	**60-**
0-2	**87.56**	10.03	1.87	0.17	0.29	0	0	0.08
4-6	26.75	**56.82**	9.19	3.87	3.05	0.32	0	0
8-13	1.02	4.49	**43.57**	8.16	35.42	6.81	0	0.53
15-20	0	2.68	15.12	**31.05**	50.01	1.01	0	0.13
25-32	0.08	1.13	2.72	7.18	**78.31**	8.58	0.73	1.27
38-43	1.05	1.8	1.86	2.1	52.58	**24.81**	11.86	3.94
48-53	0	1.84	3.68	2.27	33.33	10.62	**29.7**	18.56
60-	0	1.79	2.7	6.32	14.84	6.34	7.05	**60.96**

The performance of the Joint Fine-Tuning networks based on amplified feature sets is evaluated using the proposed robust feature sets as shown in Figure 11. In Figure 11, the two robust feature sets are fed to the network, the facial features concatenated with their robust features are fed to the first part of the network while the superpixels features and their robust features are fed to the second part of the network. Then the amplified jointly fine-tuned network for these two feature sets is trained and tested as explained in section 3.2.1.

From Tables 25 and 27, it is noticed that the overall accuracy of the amplified network for the two feature sets and their robust features outperform the accuracy result of using each feature set alone. It is also noticed that the performance of the network using the facial features concatenated with their robust features were better than the superpixels features and their robust features for some classes.

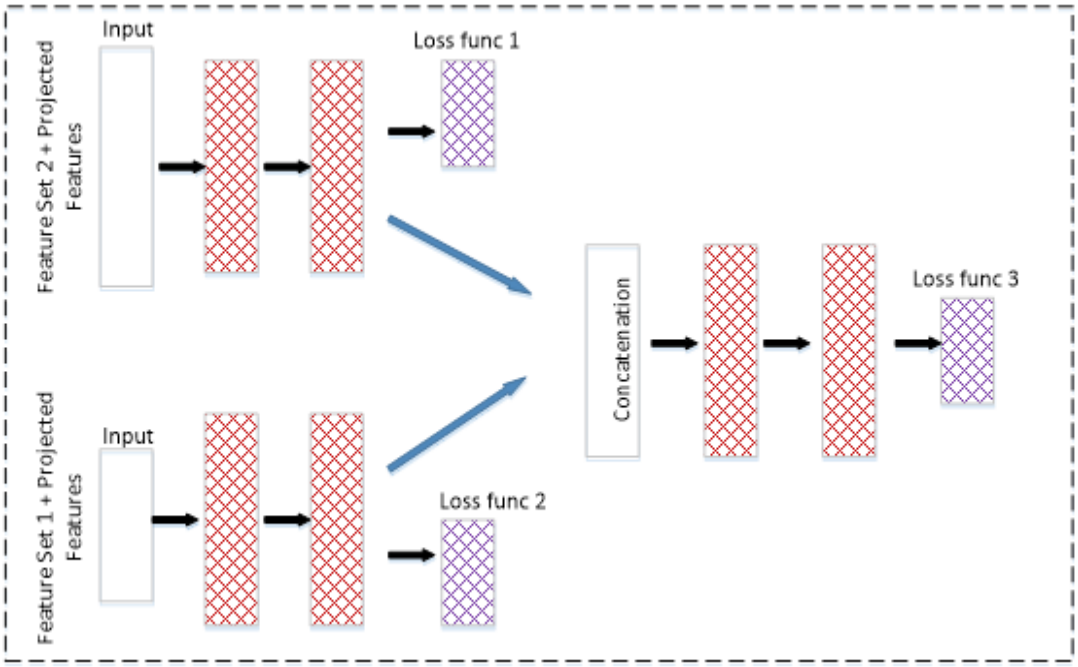

Figure 11. Jointly fine-tuned amplified network with facial and superpixels with their robust features as input.

Moreover, the performance of the Joint Fine-Tuning network based on the proposed cost function is evaluated using the proposed robust feature sets as shown in Figure 12. The two robust feature sets and their robust features are trained and tested as shown in Figure 12 and as explained in section 3.2.2.

From Tables 25, 26, and 27, it is noticed that the overall accuracy of the two jointly fine-tuned networks for the two feature sets and their robust features outperform the accuracy result of using each feature set alone. It is also noticed that the performance of the networks using the facial features concatenated with their robust features were better than the superpixels features and their robust features for some classes.

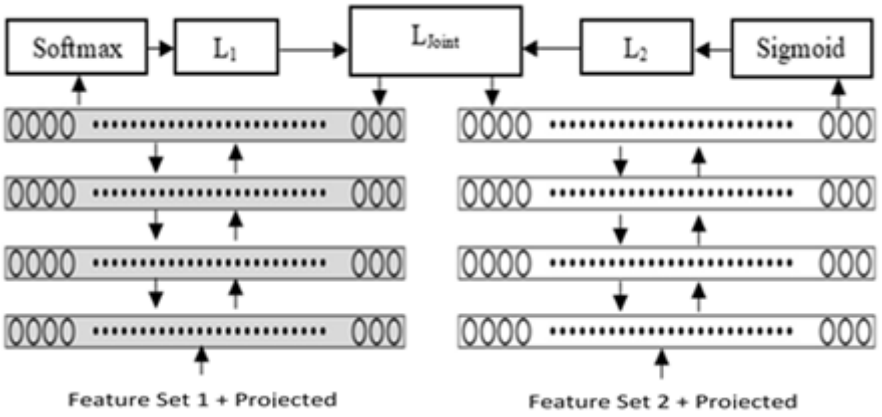

Figure 12. The proposed cost function with facial and superpixels with their robust features as input.

Table 25. Confusion matrix for the facial and superpixels features concatenated with their robust features using the jointly fine-tuned amplified network (%).

Predicted / *Actual*	**0-2**	**4-6**	**8-13**	**15-20**	**25-32**	**38-43**	**48-53**	**60-**
0-2	**88.5**	11.44	0.06	0	0	0	0	0
4-6	20.23	**63.44**	15.28	1.05	0	0	0	0
8-13	0	4.14	**49.84**	30.64	14.66	0.72	0	0
15-20	0	0.46	8.06	**37.97**	52.63	0.88	0	0
25-32	0	0.77	1.91	4.06	**82.55**	9.9	0.66	0.15
38-43	0	0.38	1.17	1.92	44.39	**43.29**	4.21	4.64
48-53	0	0	0	0.62	9.78	43.96	**27.92**	17.72
60-	0	0	0	2.14	3.19	4.25	8.81	**81.61**

Table 26. Confusion matrix for the facial and superpixels features concatenated with their robust features using the proposed cost function (%).

Predicted / *Actual*	**0-2**	**4-6**	**8-13**	**15-20**	**25-32**	**38-43**	**48-53**	**60-**
0-2	**88.72**	9.42	1.2	0	0.66	0	0	0
4-6	20.1	**68.58**	6.25	2.12	2.74	0.21	0	0
8-13	0.03	6.68	**48.23**	10.33	30.08	4.65	0	0
15-20	0	2.48	13.63	**34.86**	47.96	1.07	0	0
25-32	0	0.14	1.65	3.91	**82.56**	11.36	0.02	0.36
38-43	0.09	2.65	3.22	3.6	48.06	**35.89**	4.27	2.22
48-53	0	0.62	1.11	1.52	17.74	21.56	**28.46**	28.99
60-	0	0.35	0.52	1.92	2.34	2.44	8.76	**83.67**

Table 27. Overall classification accuracies for facial and superpixels robust features using the proposed cost function and the jointly fine-tuned amplified network on adience database (%).

	0-2	4-6	8-13	15-20	25-32	38-43	48-53	60-	Overall Accuracy	1-off Acc
Superpixels Features concatenated with their robust features	87.56	56.82	43.57	31.05	78.31	24.81	29.70	60.96	**58.31**	**86.17**
Facial Features concatenated with their robust features	86.96	65.96	45.88	35.24	78.98	41.22	15.35	83.66	**63.22**	**94.38**
Amplified Network	88.5	63.44	49.84	37.97	82.55	43.29	27.92	81.61	**65.55**	**94.86**
Jointly Tuned Network using the proposed cost function	88.72	68.58	48.23	34.86	82.56	35.89	28.46	83.67	**65.20**	**91.39**

The proposed jointly fine-tuned networks using the proposed robust features enhanced the overall accuracy for the facial age estimation as observed from Table 27. Moreover, it can be noticed that both jointly fine-tuned networks achieved considerable results with a slightly better performance for the amplified network in terms of exact accuracy. However, the 1-off accuracy results for the amplified network is much better than those for the jointly fine-tuned network using the proposed cost function. And the later results agree with the results obtained in sections 5.2 and 5.3.

5.6 Comparisons with Previous Works

The proposed architectures and models are compared with state-of-the-art results in Table 28. The three proposed methods outperform the previous state-of-the-art methods in terms of the exact and the 1-off classification accuracy. In [64] the dropout-SVM approach was used to avoid overfitting and face alignment technique was introduced to help solving the uncertainties of the facial feature extractor. In this work, the facial images are not aligned in the pre-processing phase for extracting the feature set. The work of [67] was

the first step for classifying the age and gender from face images using DNNs. [67] used a relatively simple and shallow network for feature extraction and classification, and proposed to use the over-sampling technique to partially solve the challenge of the faces misalignment in the images. Most notably, our results are significantly better than [91]. [91] collected hundreds of thousands of images to train face identification model to be used as base model for extracting image features for the proposed age classification system. In contrast, facial features extracted from a well-trained model for face recognition are used without the need for any extra data. In this work, the three proposed methods introduced solutions for enhancing age classification from different aspects. The first method focused on finding better feature sets that leads to better classification. While the second method enhanced the accuracy by jointly amplifying different feature sets. The third method enhanced the age classification by introducing an efficient new cost function which is able to estimate the error accurately. Therefore, our results were better than the results of the previous work.

Table 28. Comparison of state-of-the-art results (%).

Method		**Exact Accuracy**	**1-off Accuracy**
Previous Work	[64]	45.1	79.5
	[67] using single crop	49.5	84.6
	[67] using over-sample	50.7	84.7
	[91] chen	52.88	88.45
Our Work	**Fine-Tuned VGG-Face for Age**	**57.45**	**94.32**
	Combined-Models with dimensionality reduction	**62.26**	**92.63**
	CNN-S	**53.62**	**81.40**
	CNN-F	**57.45**	**94.32**
	CNN-FS	**63.78**	**93.70**
	DNN1	**57.45**	**94.32**
	DNN2	**53.62**	**81.40**

	Joint FineTuned with the proposed cost function	**63.78**	**93.70**
	(1) Superpixels Features concatenated with their robust features	**58.31**	**86.17**
	(2) Facial Features concatenated with their robust features	**63.22**	**94.38**
	(1) + (2) + Amplified Network	**65.55**	**94.86**
	(1) + (2) + Jointly Tuned Network using the proposed cost function	**65.20**	**91.39**

CHAPTER 6: CONCLUSIONS AND FUTURE WORK

Age estimation of the subjects from their face images is considered as an important task for many applications. Although its importance is recognized, it has received less attention than other image classification tasks. Unlike most of the previous studies that used constrained face images, a database that has unconstrained face images reflecting the variations of the real subject images taken from the internet repositories is used.

In this work, different methods are proposed to enhance the age classification from unconstrained face images. First, it is investigated the employment of CNNs, which were previously trained for different tasks on large databases, in the design of a DNN for age classification task. Existing benchmarks for age classifications are relatively small compared to the benchmarks used in face recognition. Training a deep neural architecture using a small benchmark is problematic since training a very deep CNN architecture on a relatively small benchmark is liable to a critical overfitting. To overcome this problem, we use deep CNN architectures that are trained for other classification tasks such as image classification, semantic segmentation, and face recognition on large benchmarks. Then, these architectures are adapted and fine-tuned to estimate the age of a subject from an unconstrained 2D image. A deep pre-trained CNN model for face recognition is used to extract facial features. Then, these extracted facial features are used to train a DNN for age classification. In addition, dimensionality reduction is performed on the last convolutional layer features of these pre-trained models. The dimensionally reduced features are then incorporated and trained to estimate the age by using DNN architecture. Despite the difficulty of building a large unconstrained real-word benchmark containing millions of

face images from social media websites and internet repositories for age estimation, we hope that such benchmarks will become available in near future. The availability of large benchmarks will help improving the current results further especially by using very deep CNNs that have shown remarkable performances in other classification fields.

Second, a new cost function is proposed for jointly fine-tuning two DDNs. Two DNNs are trained and tuned concurrently on different feature sets that are extracted from the same training sample. The first feature set contained facial features from a pre-trained model for face recognition. The second feature set is extracted by dividing the image into homogeneous superpixels and then finding information from these superpixels and their relations with adjacent superpixels. The output of the last convolutional layer of the pre-trained models is used for feature extraction. The high dimensional feature vector that is obtained from each training image is reduced by using the PCA. One of the benefits of the proposed cost function is the ability to reduce the effect of the overfitting problem. This is achieved by involving the propagated errors generated from two networks that perform simultaneous learning process on two feature sets. The two networks calculated the error depending on two different cost functions, where each one has a different approach to calculate the error. Another advantage of the proposed work is to involve different feature sets in order to optimize the network parameters and to minimize the overfitting problem further. Moreover, the jointly fine-tuning of two networks provides a platform to combine and extract the distinctive features of two different feature spaces by coupling the learning process of two networks using the proposed cost function. Using a fixed and a unified learning rate for the jointly fine-tuned networks based on the proposed cost function leads to propagate different and incompatible error rates without reflecting the actual joint

learning. The later happens when the used feature sets are loosely related and each set requires different network parameters in order to extract the desired patterns of data that contains higher representations than the initial form of the feature sets. Different learning rates have been calibrated automatically for the jointly fine-tuned network to reach stable learning ratios between the two DNNs. Calibrating the learning rates of the two DNNs was quite fast and non-problematic, hence the jointly fine-tuned networks converged in a reasonable time. A possible implication is that the number of networks and feature sets will affect the performance of the proposed jointly fine-tuned networks. For instance, if three networks with three feature sets are used, the cost function should be modified in order to calculate the effect of the third feature set especially if the new set is unrelated to the other two feature sets.

Finally, we propose a new model based on convolutional neural networks (CNNs) and $l2$,1-norm to select age-related features for the age estimation task. A new cost function is proposed. To learn and train the new model, we provide the analysis and the proof for the convergence of the new cost function to solve the minimization problem of deep neural networks (DNNs) and the $l2$,1-norm. High-level features are extracted from the facial images by using transfer learning. Then, the extracted features are fed to the proposed model to select the most efficient age-related features (the robust features). The robust features achieved classification accuracies that are comparable to the performance of the original features. It verifies the effectiveness of the proposed model in finding robust features for the age information from the input image. Moreover, the small dimension of the robust feature set requires a relatively smaller network for training and it reduces the computational time. The proposed framework based the new cost function combines the

efficiency and powerfulness of DNNs for extracting distinctive features and the l_{21}-*norm* for selecting robust features by reducing the effect of the outliers. The l_{21}-*norm* is well known for its ability to deal with the outliers in facial images. Since unconstrained images contains a variety of outlier images, the usage of the l_{21}-*norm* based sigmoid cost function allowed our model to focus on finding robust age-related features other than focusing on outliers.

We provide extensive experimental results on a public database, which demonstrate the capability of our proposed work to classify the age from the facial images. And it is shown that the proposed methods for age classification from unconstrained facial image outperformed stat-of-the-art results.

REFERENCES

[1] H. Han, C. Otto, and A. K. Jain, "Age estimation from face images: Human vs. machine performance,"in *International Conference on Biometrics (ICB)*, 2013, pp. 1-8.

[2] W.-L. Chao, J.-Z. Liu, and J.-J. Ding, "Facial age estimation based on label-sensitive learning and age-oriented regression," Pattern Recognition, vol. 46(3), pp. 628-641, 2013.

[3] J. Galbally, S. Marcel, and J. Fierrez, "Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition," *IEEE transactions on image processing,* vol. 23, pp. 710-724, 2014.

[4] J. Galbally, S. Marcel, and J. Fierrez, "Biometric antispoofing methods: A survey in face recognition," *IEEE Access,* vol. 2, pp. 1530-1552, 2014.

[5] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial Intelligence Review,* vol. 43, pp. 1-54, 2015.

[6] Y. Xu, X. Fang, X. Li, J. Yang, J. You, H. Liu*, et al.*, "Data uncertainty in face recognition," *IEEE transactions on cybernetics,* vol. 44, pp. 1950-1961, 2014.

[7] Y. Xu, X. Li, J. Yang, Z. Lai, and D. Zhang, "Integrating conventional and inverse representation for face recognition," *IEEE transactions on cybernetics,* vol. 44, pp. 1738-1746, 2014.

[8] J. Lu, V. E. Liong, X. Zhou, and J. Zhou, "Learning compact binary face descriptor for face recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol. 37, pp. 2041-2056, 2015.

[9] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum, "Deep convolutional inverse graphics network," in *Advances in Neural Information Processing Systems*, 2015, pp. 2539-2547.

[10] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics,* vol. 20, pp. 413-425, 2014.

[11] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR),* 2005, pp. 539-546.

[12] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *IEEE 12th International Conference on Computer Vision,* 2009, pp. 365-372.

[13] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701-1708.

[14] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proceedings of the Second IEEE Workshop on Applications of Computer Vision,* 1994, pp. 138-142.

[15] F. Samaria and S. Young, "HMM-based architecture for face identification," *Image and vision computing,* vol. 12, pp. 537-543, 1994.

[16] V. Blanz, S. Romdhani, and T. Vetter, "Face identification across different poses and illuminations with a 3d morphable model," in *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition,* 2002, pp. 202-207.

[17] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience,* vol. 3, pp. 71-86, 1991.

[18] M. S. Bartlett, "Independent component representations for face recognition," in *Face Image Analysis by Unsupervised Learning*, 2001, pp. 39-67.

[19] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 19, pp. 711-720, 1997.

[20] D. L. Swets and J. J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 18, pp. 831-836, 1996.

[21] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural computation,* vol. 10, pp. 1299-1319, 1998.

[22] M.-H. Yang, "Kernel Eigenfaces vs. Kernel Fisherfaces: Face Recognition Using Kernel Methods," in Proceedings of the Fifth IEEE International Conference on automatic face and gesture recognition, pp. 215-220, 2002.

[23] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science,* vol. 290, pp. 2323-2326, 2000.

[24] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815-823.

[25] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," in *Proceedings of the British Machine Vision* 2015, pp. 1(3), 6.

[26] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5325-5334.

[27] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 2015, pp. 643-650.

[28] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in neural information processing systems*, 2014, pp. 1988-1996.

[29] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891-1898.

[30] Y. Tian, T. Kanade, and J. F. Cohn, "Facial expression recognition," in *Handbook of face recognition*, 2011, pp. 487-519.

[31] P. K. Manglik, U. Misra, Prashant, and H. B. Maringanti, "Facial expression recognition," in *IEEE International Conference on Systems, Man and Cybernetics*, 2004, pp. 2220-2224.

[32] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing,* vol. 27, pp. 803-816, 2009.

[33] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition,* 1998, pp. 454-459.

[34] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision and image understanding,* vol. 91, pp. 160-187, 2003.

[35] M. Pantic, "Facial expression recognition," in *Encyclopedia of biometrics*, ed: Springer, 2009, pp. 400-406.

[36] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE transactions on image processing,* vol. 16, pp. 172-187, 2007.

[37] C. Shan, S. Gong, and P. W. McOwan, "Robust facial expression recognition using local binary patterns," *in IEEE International Conference on Image Processing (ICIP)*, 2005, pp. 914-917.

[38] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of ACM on International Conference on Multimodal Interaction*, 2015, pp. 435-442.

[39] S. K. D'mello and A. Graesser, "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features," *User Modeling and User-Adapted Interaction,* vol. 20, pp. 147-187, 2010.

[40] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz*, et al.*, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine,* vol. 18, pp. 32-80, 2001.

[41] I. Cohen, A. Garg, and T. S. Huang, "Emotion recognition from facial expressions using multilevel HMM," in *Neural information processing systems (NIPS) on Workshop Affective Computation*, 2000.

[42] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern recognition,* vol. 36, pp. 259-275, 2003.

[43] L. C. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information," in *Proceedings of International Conference on Information, Communications and Signal Processing (ICICS),* 1997, pp. 397-401.

[44] R. Adolphs, H. Damasio, D. Tranel, and A. R. Damasio, "Cortical systems for the recognition of emotion in facial expressions," *Journal of neuroscience,* vol. 16, pp. 7678-7687, 1996.

[45] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1805-1812.

[46] R. Valenti, N. Sebe, and T. Gevers, "Combining head pose and eye location information for gaze estimation," *IEEE Transactions on Image Processing,* vol. 21, pp. 802-815, 2012.

[47] S. Asteriadis, P. Tzouveli, K. Karpouzis, and S. Kollias, "Estimation of behavioral user state based on eye gaze and head pose—application in an e-learning environment," *Multimedia Tools and Applications,* vol. 41, pp. 469-493, 2009.

[48] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE transactions on pattern analysis and machine intelligence,* vol. 31, pp. 607-626, 2009.

[49] Q. X. Nguyen and S. Jo, "Electric wheelchair control using head pose free eye-gaze tracker," *Electronics Letters,* vol. 48, pp. 750-752, 2012.

[50] C. H. Morimoto and M. R. Mimica, "Eye gaze tracking techniques for interactive applications," *Computer vision and image understanding,* vol. 98, pp. 4-24, 2005.

[51] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Head pose-free appearance-based gaze sensing via eye image synthesis," in *21st International Conference on Pattern Recognition (ICPR),* 2012, pp. 1008-1011.

[52] P. M. Corcoran, F. Nanu, S. Petrescu, and P. Bigioi, "Real-time eye gaze tracking for gaming design and consumer electronics systems," *IEEE Transactions on Consumer Electronics,* vol. 58, pp. 347-355, 2012.

[53] F. Lu, T. Okabe, Y. Sugano, and Y. Sato, "Learning gaze biases with head motion for head pose-free gaze estimation," *Image and Vision Computing,* vol. 32, pp. 169-179, 2014.

[54] C. B. Ng, Y. H. Tay, and B.-M. Goi, "Recognizing human gender in computer vision: a survey," in *Pacific Rim International Conference on Artificial Intelligence*, 2012, pp. 335-346.

[55] T. Huynh, R. Min, and J.-L. Dugelay, "An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data," in *Asian Conference on Computer Vision*, 2012, pp. 133-145.

[56] B. Yang and S. Chen, "A comparative study on local binary pattern (LBP) based face recognition: LBP histogram versus LBP image," *Neurocomputing,* vol. 120, pp. 365-379, 2013.

[57] C. Shan, "Learning local binary patterns for gender classification on real-world face images," *Pattern Recognition Letters,* vol. 33, pp. 431-437, 2012.

[58] J. Bekios-Calfa, J. M. Buenaposada, and L. Baumela, "Robust gender recognition by exploiting facial attributes dependencies," *Pattern Recognition Letters,* vol. 36, pp. 228-234, 2014.

[59] L. Ballihi, B. B. Amor, M. Daoudi, A. Srivastava, and D. Aboutajdine, "Boosting 3-D-geometric features for efficient face recognition and gender classification," *IEEE Transactions on Information Forensics and Security,* vol. 7, pp. 1766-1779, 2012.

[60] J. Mansanet, A. Albiol, and R. Paredes, "Local deep neural networks for gender recognition," *Pattern Recognition Letters,* vol. 70, pp. 80-86, 2016.

[61] P. Rai and P. Khanna, "Gender classification techniques: A review," in *Advances in Computer Science, Engineering & Applications*, 2012, pp. 51-59.

[62] G. Antipov, S.-A. Berrani, and J.-L. Dugelay, "Minimalistic CNN-based ensemble model for gender prediction from face images," *Pattern recognition letters,* vol. 70, pp. 59-65, 2016.

[63] J. van de Wolfshaar, M. F. Karaaba, and M. A. Wiering, "Deep convolutional neural networks and support vector machines for gender recognition," in *IEEE Symposium Series on Computational Intelligence,* 2015, pp. 188-195.

[64] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," IEEE Transactions on Information Forensics and Security, vol. 9(12), pp. 2170-2179, 2014.

[65] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," IEEE transactions on pattern analysis and machine intelligence, vol. 32(11), pp. 1955-1976, 2010.

[66] A. C. Gallagher and T. Chen, "Understanding images of groups of people,"in IEEE Conference on Computer Vision and Pattern Recognition. (CVPR), 2009, pp. 256-263.

[67] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 34-42.

[68] Y. H. Kwon and N. da Vitoria Lobo, "Age classification from facial images," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Proceedings (CVPR), 1994, pp. 762-767.

[69] L. G. Farkas, "Anthropometry of the Head and Face," Raven Press, New York, 1994.

[70] J. Hayashi, M. Yasumoto, H. Ito, Y. Niwa, and H. Koshimizu, "Age and gender estimation from facial image processing," in Proceedings of the 41st SICE Annual Conference, 2002, pp. 13-18.

[71] J. Hayashi, M. Yasumoto, H. Ito, and H. Koshimizu, "Method for estimating and modeling age and gender using facial image processing," in Proceedings of Seventh International Conference on Virtual Systems and Multimedia, 2001, pp. 439-448.

[72] N. Ramanathan and R. Chellappa, "Modeling age progression in young faces," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2006, pp. 387-394.

[73] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24(4), pp. 442-455, 2002.

[74] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 29(12), pp. 2234-2240, 2007.

[75] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai, "Learning from facial aging patterns for automatic age estimation," in Proceedings of the 14th ACM international conference on Multimedia, 2006, pp. 307-316.

[76] Y. Fu, Y. Xu, and T. S. Huang, "Estimating human age by manifold analysis of face pictures and regression on aging features," in IEEE International Conference on Multimedia and Expo, 2007, pp. 1383-1386.

[77] K. Scherbaum, M. Sunkel, H. P. Seidel, and V. Blanz, "Prediction of Individual Non-Linear Aging Trajectories of Faces," in Computer Graphics Forum, 2007, pp. 285-294.

[78] Y. Fu, and T. S. Huang, "Human age estimation with regression on discriminative aging manifold," IEEE Transactions on Multimedia, vol. 10(4), pp. 578-584, 2008.

[79] A. Gunay and V. V. Nabiyev, "Automatic age classification with LBP," in 23rd International Symposium on Computer and Information Sciences (ISCIS), 2008, pp. 1-4.

[80] F. Gao and H. Ai, "Face age classification on consumer images with gabor feature and fuzzy lda method," in International Conference on Biometrics, 2009, pp. 132-141.

[81] S. Yan, M. Liu, and T. S. Huang, "Extracting age information from local spatially flexible patches," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, pp. 737-740.

[82] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. S. Huang, "Regression from patch-kernel," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1-8.

[83] J. Suo, T. Wu, S. Zhu, S. Shan, X. Chen, and W. Gao, "Design sparse features for age estimation using hierarchical face model," in 8th IEEE International Conference on Automatic Face & Gesture Recognition (FG), 2008, pp. 1-6.

[84] G. Mu, G. Guo, Y. Fu, and T. S. Huang, "Human age estimation using bio-inspired features," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 112-119.

[85] G. Guo, G. Mu, Y. Fu, C. R. Dyer, and T. S. Huang, "A study on automatic age estimation using a large database," in ICCV, 2009, pp. 1986-1991.

[86] C. Shan, "Learning local features for age estimation on real-life faces," in Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis, 2010, pp. 23-28.

[87] F. Alnajar, C. Shan, T. Gevers, and J.-M. Geusebroek, “Learning-based encoding with soft assignment for age estimation under unconstrained imaging conditions,” Image and Vision Computing, vol. 30(12), pp. 946-953, 2012.

[88] G. Guo and G. Mu, "A framework for joint estimation of age, gender and ethnicity on a large database," Image and Vision Computing, vol. 32(10), pp. 761-770, 2014.

[89] J. Lu, V. E. Liong, and J. Zhou, "Cost-sensitive local binary feature learning for facial age estimation," IEEE Transactions on Image Processing, vol. 24(12), pp. 5356-5368, 2015.

[90] R. Ranjan, S. Zhou, J. Cheng Chen, A. Kumar, A. Alavi, V. M. Patel, et al., "Unconstrained age estimation with deep convolutional neural networks," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 109-117.

[91] J.-C. Chen, A. Kumar, R. Ranjan, V. M. Patel, A. Alavi, and R. Chellappa, "A cascaded convolutional neural network for age estimation of unconstrained faces," in IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), 2016, pp. 1-8.

[92] J.-C. Chen, V. M. Patel, and R. Chellappa, "Unconstrained face verification using deep cnn features," in IEEE Winter Conference on Applications of Computer Vision (WACV), 2016, pp. 1-9.

[93] I. Huerta, C. Fernández, C. Segura, J. Hernando, and A. Prati, "A deep analysis on age estimation," Pattern Recognition Letters, vol. 68, pp. 239-249, 2015.

[94] H.-F. Yang, B.-Y. Lin, K.-Y. Chang, and C.-S. Chen, "Automatic Age Estimation from Face Images via Deep Ranking," in BMVC, 2015, pp. 55.

[95] D. Yi, Z. Lei, S.Z. Li, "Age estimation by multi-scale convolutional network, " in Proceedings of the Asian Conference on Computer Vision, 2014, pp. 144–158.

[96] X. Yang, B. Gao, C. Xing, Z. Huo, X. Wei, Y. Zhou, J. Wu, X. Geng, "Deep label distribution learning for apparent age estimation, " in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 344–350.

[97] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, X. Chen, "Agenet: Deeply learned regressor and classifier for robust apparent age estimation, " in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 258–266.

[98] R. Rothe, R. Timofte, and L. Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," International Journal of Computer Vision, 2016, pp. 1-14.

[99] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in British Machine Vision Conference, 2015, p. 6.

[100] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” Journal of Machine Learning Research, vol. 15(1), pp. 1929-1958, 2014.

[101] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al., "Going deeper with convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1-9.

[102] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778.

[103] K. Simonyan, and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in ICLR, 2015.

[104] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431-3440.

[105] L. Cao, K. S. Chua, W. Chong, H. Lee, and Q. Gu, “A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine,” Neurocomputing, vol. 55(1), pp. 321-336, 2003.

[106] A. Ghodsi, “Dimensionality reduction a short tutorial,” Department of Statistics and Actuarial Science, Univ. of Waterloo, Ontario, Canada, vol. 37, pp. 38, 2006.

[107] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," in Proceedings of the 12th IAPR International Conference on Pattern Recognition, A: Computer Vision & Image Processing, 1994, pp. 582-585.

[108] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5162-5170.

[109] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, pp. 1929-1958, 2014.

[110] Q. Gu, Z. Li, and J. Han, "Joint feature selection and subspace learning," in Proceedings-International Joint Conference on Artificial Intelligence (IJCAI), 2011, pp. 1294-1299.

[111] R. He, T. N. Tan, L. Wang, and W. Zheng, "L21 regularized correntropy for robust feature selection," In CVPR, 2012, pp. 2504–2511.

[112] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selction via joint L21-norms minimization," In NIPS, 2010, pp. 1813–1821.

[113] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," International Journal of Computer Vision, vol. 115(3), pp. 211-252, 2015.

[114] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," International journal of computer vision, vol. 88(2), pp. 303-338, 2010.